DATA MINING AND STATISTICAL METHODS FOR KNOWLEDGE DISCOVERY IN DISEASES BASED ON MULTIMODAL OMICS

EDITED BY: Jiajie Peng, Tao Wang and Miguel E. Renteria

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-174-6
DOI 10.3389/978-2-88976-174-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

DATA MINING AND STATISTICAL METHODS FOR KNOWLEDGE DISCOVERY IN DISEASES BASED ON MULTIMODAL OMICS

Topic Editors:
Jiajie Peng, Northwestern Polytechnical University, China
Tao Wang, Northwestern Polytechnical University, China
Miguel E. Renteria, The University of Queensland, Australia

Citation: Peng, J., Wang, T., Renteria, M. E., eds. (2022). Data Mining and Statistical Methods for Knowledge Discovery in Diseases Based on Multimodal Omics. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-174-6

Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 26 April 2022
doi: 10.3389/fgene.2022.895796

Editorial: Data Mining and Statistical Methods for Knowledge Discovery in Diseases Based on Multimodal Omics

Tao Wang[1,2]*, *Miguel E. Rentería*[3]* and *Jiajie Peng*[1,2]*

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China, [2]Key Laboratory of Big Data Storage and Management, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an, China, [3]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

Keywords: multimodal, omics, disease biology, data mining, statistical methods

Editorial on the Research Topic

Data Mining and Statistical Methods for Knowledge Discovery in Diseases Based on Multimodal Omics

Over the last decade, advances in high-throughput omics technologies and methods have enabled researchers to measure multiple biological data modalities simultaneously and accurately or to integrate multi-omics data from different sources and modalities. Numerous datasets are being rapidly generated encompassing genomics, transcriptomics, proteomics, metabolomics, phenomics, radiomics, cutting-edge 3D spatial omics, and single-cell omics data. This represents an unprecedented opportunity for knowledge discovery in disease biology, including the identification of biomarkers, functional modules, causal pathways, or regulatory networks implicated in disease, thus having also the potential to bolster current therapeutic pipelines.

In parallel, a wide-array of statistical methods have been developed to leverage availability of these data, from genome-wide association studies (GWAS) to transcription-wide association studies (TWAS), methylome-wide association studies (MWAS), molecular quantitative trait loci (molQTL) analysis, or summary-based two-sample Mendelian Randomization. However, the ability to integrate different features of existing methods is still insufficient, limiting the power for knowledge discovery. Thus, advances in data mining, or statistical and machine learning techniques are urgently needed to perform cross-modal data integration and modeling. Here, we present a Research Topic on "Data Mining and Statistical Methods for Knowledge Discovery in Diseases Based on Multimodal Omics" to showcase studies that leverage these techniques to enable discovery of disease-related knowledge and illuminate molecular mechanisms of complex diseases. After rigorous peer-review, a total of 14 outstanding articles were selected for this topic collection. Below we highlighted six of them.

Huang et al. explored the causal effects of insomnia on bipolar disorder, major depression, and schizophrenia in the European population using a two-sample Mendelian randomization approach. They first collected GWAS summary datasets for each trait and conducted meta-analyses for each trait to increase statistical power. The results of Mendelian randomization were further evaluated using extensive complementarity and sensitivity analysis. Among these psychiatric disorders, they found insomnia is causally associated with an increased risk of major depression, with an odds ratio estimated as 1.408 (95% confidence interval (CI): 1.210–1.640, p = 1.03E-05) in the European population. No causal association was observed for other traits. The study provides new evidence to support the causal effect of insomnia on major depression and adds to a better understanding of the relationship between sleep and psychiatric disorders.

OPEN ACCESS

Edited and reviewed by:
Simon Charles Heath, Center for Genomic Regulation (CRG), Spain

***Correspondence:**
Tao Wang
twang@nwpu.edu.cn
Miguel E. Rentería
Miguel.Renteria@qimrberghofer.edu.au
Jiajie Peng
jiajiepeng@nwpu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *14 March 2022*
Accepted: *25 March 2022*
Published: *26 April 2022*

Citation:
Wang T, Rentería ME and Peng J (2022) Editorial: Data Mining and Statistical Methods for Knowledge Discovery in Diseases Based on Multimodal Omics. Front. Genet. 13:895796. doi: 10.3389/fgene.2022.895796

Hamidi et al. proposed a machine learning framework to explore miRNA biomarkers and prediction for Ovarian cancer. miRNAs play an important role in cancer progression. In this study, the authors first used LASSO and Elastic Net for miRNA feature selection. They found 10 miRNA's as potential biomarkers by comparing the expression levels in ovarian serum cancer samples and normal samples. Furthermore, they used multiple machine learning classifiers, including logistic regression, random forest, artificial neural network, XGBoost, and decision trees for ovarian cancer prediction. Experiments demonstrated the accuracy of their proposed model. The performance of the proposed models was further evaluated in external datasets.

Cerebral ischemic stroke (IS) is a complex disease caused by multiple factors, including vascular risk, genetic, and environmental factors. Identifying the genes associated with IS critical for understanding the biological mechanisms underlying the disease. Liu et al. proposed a network representation learning (NRL)-based method to identify the disease-related genes of cerebral IS. The proposed method includes three key components: capturing the topological information of the PPI network, denoising the gene feature, and optimizing a support vector machine (SVM) classifier to identify IS-related genes. The evaluation showed that the proposed method performs better than existing methods on IS-related gene prediction. In addition, the case study also shows that the proposed method can identify IS-related genes.

Recently, single-cell RNA sequencing (scRNA-seq) technology has been used to measure RNA levels at single-cell resolution to study biological functions. Xu et al. proposed an imputation method based on semi-supervised autoencoders named AdImpute. The method applies the cost function with imputation weights to learn the latent information in the data to achieve a more accurate imputation. The evaluation indicates that AdImpute is more accurate than the other four publicly available scRNA-seq imputation methods on the simulated and real data sets.

Yang et al. tackled the issue of systematic selection bias in Mendelian randomization. The authors proposed a new approach that uses control exposures based on subject-matter knowledge to triangulate the estimated causal effects vulnerable to selection bias. The proposed approach can be used to assess credible MR estimates in the presence of selection bias from selection of survivors. The authors illustrate the application of their method by validating MR estimates through a real example investigating the potential association of transferrin with stroke (including ischemic and cardioembolic stroke).

Park et al. developed an innovative approach for integrative pathway analysis that leverages genome-wide association studies summary statistics to construct genetic metabolomic scores (GMSs) that are then used as components of pathways in a hierarchical model that considers the structural relationships of SNPs, metabolites, pathways, and phenotypes. The authors applied their method to identify pathways associated with type 2 diabetes in the Korean population.

All the contributions in this special issue have been peer-reviewed by no less than two professional domain experts. We believe that the final compilation includes high-quality publications that represent significant scientific progress that will impact the relevant research communities. On this basis, we have launched a second edition of this Research Topic which is currently open for submissions.

AUTHOR CONTRIBUTIONS

TW, MR and JP conducted this topic issue and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (Nos. 62102319 and 62072376); Fundamental Research Funds for the Central Universities of China (No. G2021KY05112).

ACKNOWLEDGMENTS

We would like to thank all authors for their contributions to our special issue and all reviewers' time and effort. We would also thank the editor-in-chief and editorial department of Frontiers in Genetics for their support throughout the process.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fgene.2021.698124

OPEN ACCESS

Edited by:
Tao Wang,
Northwestern Polytechnical University, China

Reviewed by:
Dongjin Wang,
Nanjing Drum Tower Hospital, China
Jing Zhang,
Shanghai Jiao Tong University, China
Jiaqiang Zhu,
University of Michigan, United States

***Correspondence:**
Yi-Ning Yang
yangyn5126@163.com
Di Liu
liud@wh.iov.cn

†*These authors have contributed equally to this work*

‡**ORCID:**
Yan-Hong Li
orcid.org/0000-0002-5449-729X
Ying Cao
orcid.org/0000-0002-4289-452X
Fen Liu
orcid.org/0000-0002-8696-303X
Qian Zhao
orcid.org/0000-0002-8770-3838
Dilare Adi
orcid.org/0000-0002-2480-4703
Qiang Huo
orcid.org/0000-0003-2480-660X
Zheng Liu
orcid.org/0000-0002-2762-2163
Jun-Yi Luo
orcid.org/0000-0002-4767-1288
Bin-Bin Fang
orcid.org/0000-0002-9072-5106
Ting Tian
orcid.org/0000-0002-1332-5424
Di Liu
orcid.org/0000-0003-3693-2726
Xiao-Mei Li
orcid.org/0000-0003-4152-8541
Yi-Ning Yang
orcid.org/0000-0002-8332-8508

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *20 April 2021*
Accepted: *07 June 2021*
Published: *28 June 2021*

Visualization and Analysis of Gene Expression in Stanford Type A Aortic Dissection Tissue Section by Spatial Transcriptomics

Yan-Hong Li[1,2,3†‡], Ying Cao[4,5,6,7†‡], Fen Liu[3,8†‡], Qian Zhao[1,3‡], Dilare Adi[3,8‡], Qiang Huo[9‡], Zheng Liu[9‡], Jun-Yi Luo[1,3‡], Bin-Bin Fang[3,8‡], Ting Tian[3‡], Xiao-Mei Li[1,3‡], Di Liu[1,4,6,10*‡] and Yi-Ning Yang[1,3,11*‡]

[1] Department of Cardiology, First Affiliated Hospital of Xinjiang Medical University, Urumqi, China, [2] Department of Clinical Laboratory, First Affiliated Hospital of Xinjiang Medical University, Urumqi, China, [3] State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asian, Department of Cardiology, First Affiliated Hospital of Xinjiang Medical University, Urumqi, China, [4] Computational Virology Group, Center for Bacteria and Virus Resources and Application, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China, [5] CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China, [6] University of Chinese Academy of Sciences, Beijing, China, [7] College of Animal Sciences and Veterinary Medicine, Guangxi University, Nanning, China, [8] State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, First Affiliated Hospital of Xinjiang Medical University, Urumqi, China, [9] Department of Cardiac Surgery, First Affiliated Hospital of Xinjiang Medical University, Urumqi, China, [10] Xinjiang Medical University, Urumqi, China, [11] People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, China

Background: Spatial transcriptomics enables gene expression events to be pinpointed to a specific location in biological tissues. We developed a molecular approach for low-cell and high-fiber Stanford type A aortic dissection and preliminarily explored and visualized the heterogeneity of ascending aortic types and mapping cell-type-specific gene expression to specific anatomical domains.

Methods: We collected aortic samples from 15 patients with Stanford type A aortic dissection and a case of ascending aorta was randomly selected followed by 10x Genomics and spatial transcriptomics sequencing. In data processing of normalization, component analysis and dimensionality reduction analysis, different algorithms were compared to establish the pipeline suitable for human aortic tissue.

Results: We identified 19,879 genes based on the count level of gene expression at different locations and they were divided into seven groups based on gene expression trends. Major cell that the population may contain are indicated, and we can find different main distribution of different cell types, among which the tearing sites were mainly macrophages and stem cells. The gene expression of these different locations and the cell types they may contain are correlated and discussed in terms of their involvement in immunity, regulation of oxygen homeostasis, regulation of cell structure and basic function.

Conclusion: This approach provides a spatially resolved transcriptome– and tissue-wide perspective of the adult human aorta and will allow the application of human fibrous

aortic tissues without any effect on genes in different layers with low RNA expression levels. Our findings will pave the way toward both a better understanding of Stanford type A aortic dissection pathogenesis and heterogeneity and the implementation of more effective personalized therapeutic approaches.

Keywords: spatial transcriptomics, aortic, Stanford type A aortic dissection, gene expression, bioinformatics

INTRODUCTION

Stanford type A aortic dissection (AAD) is the most common thoracic aortic disease, which has a high degree of morbidity and leads to extensive medical expenditure for survivors. It may rapidly fatal if not diagnosed early and managed appropriately (Guo et al., 2016). From the biomechanical viewpoint, the mechanism of injury is based on the inability of the vascular wall to withstand high shear stress that penetrates the intimal vessel layer, resulting in blood flow to the intimal and medial layers or disruption of the media layer (Yang et al., 2020). The pathological features of aortic tissue are characterized by an enlarged and degenerative medial layer, loss or dysfunction of vascular smooth muscle cells (VSMCs), proteoglycan accumulation, and collagen and elastic fiber cross-linked disorder and fragmentation (Oller et al., 2017). The risk factors associated with the occurrence and development of AAD include hypertension, dyslipidemia, atherosclerosis, cigarette smoking, and male gender (Yang et al., 2020). Although the major aortic cell types in the whole aorta are well known (Dobnikar et al., 2018; Kalluri et al., 2019), the heterogeneity and relative contribution of different vascular cells in AAD are poorly understood.

Previous studies have demonstrated the reason for the tear during aortic dissection by regular transcriptome analyses off multiple pathways (Huang et al., 2018; Yang et al., 2018; Wang et al., 2019). However, these studies have certain limitations in accuracy. Based on single cell RNA sequencing (scRNA-seq), we can only determine the average gene expression of the ruptured tissue, and many details are lost. Identification of aortic cell-type composition depends on anatomy, and methods such as radiography and pathology may affect the reliability of the results. How the cells of the intima, media, and adventitia are affected and the role they play in the occurrence and development of the disease need to be studied urgently using novel techniques.

Spatial transcriptomics (ST) is an approach that allows the visualization and quantitative analysis of the transcriptome with spatial resolution in individual tissue samples (Stahl et al., 2016). By placing tissue sections on glass slides with arrayed oligonucleotides containing positional barcodes, high-quality cDNA libraries can be generated with precise positional information for RNA sequencing. ST has been used to study the mouse olfactory bulb (Stahl et al., 2016), breast cancer (He et al., 2020), adult human heart tissues (Asp and Salmen, 2017), melanoma tissues (Thrane et al., 2018), prostate cancer tissues (Berglund et al., 2018), gingival tissues (Lundmark et al., 2018), mouse and human spinal cord tissues (Maniatis et al., 2019), and model plant species (Giacomello et al., 2017). Asp et al. (2019) used ST to reveal the comprehensive transcriptional landscape of cell types populating the embryonic heart at three developmental stages and mapped cell type-specific gene expression to specific anatomical domains. They identified unique gene profiles that corresponded to distinct anatomical regions in each developmental stage using ST (Asp et al., 2019). High-resolution spatial heterogeneity can be captured, and the rich spatial information regarding unbiased gene expression for cells and tissues can be retained in ST results, compared with results of regular transcriptome analyses using bulk sequencing or scRNA-seq (Gerlinger et al., 2012). The heterogeneity of gene expression and spatial organization in the aorta may help identify the underlying pathogenesis of aortic dissection. Considering the particularity of aortic structure and cell composition, there is still no suitable bioinformatics algorithm to analyze the ST sequencing data of the aorta.

Here, for the first time, we provided an algorithm suitable for aortic tissue and analyzed aortic tears simultaneously at the tissue- and transcriptome-wide scales using the ST, which allowed for the identification and spatial mapping of distinct cell types, subpopulations, and cell states within heterogeneous samples. We identified the top 20 spatially related genes and identified major cell types, including smooth muscle cells (SMCs), fibroblasts, endothelial cells (ECs), and infiltrated immune cells (including macrophages, B cells, T cells, and dendritic cells). Further analysis showed that different types of cells showed different enrichment of signal pathways. For example, cluster M4 was mainly composed of macrophages and Kupffer cells, and the signaling pathways were mainly related to immunity and apoptosis. The establishment of these profiles is the first step toward obtaining an unbiased view of aortic dissection and can serve as a reference for future studies on AAD.

MATERIALS AND METHODS

Ethics

AAD participants gave written informed consent, permission for tissue analyses, and consent for the collection of relevant clinical data before enrolling in the study as approved by the ethics committee at the First Affiliated Hospital of Xinjiang Medical University (Urumqi, China) (20150006-8). All procedures were conformed to the principles outlined in the Declaration of Helsinki.

Participants

We recruited 15 adult patients with AAD (along with their demographic information such as age, sex, etc.) admitted to the First Affiliated Hospital of Xinjiang Medical University (Urumqi, China) from September 1, 2019, to July 1, 2020. The patients were diagnosed through history, findings of physical

examination, and imaging findings according to currently accepted standards (Erbel et al., 2014). Patients were excluded if they had Marfan syndrome, Ehlers Danlos syndrome, Loeys-Dietz syndrome, Turner syndrome, congenital bi-leaflet aortic valves, aortic aneurysm, traumatic dissection, or other connective tissue disorders, and those aged < 18 years.

Collection and Preparation of Aortic Tissue

The aortic sample was rapidly collected within 30 min after excision. The specimen was rinsed at least five times in precooled saline; then, the thrombus and redundant tissues were removed immediately using eye scissors and sterile tweezers on a clean Petri dish (placed on dry ice). Tissue samples were then sliced into approximately 6 × 6 mm sections, embedded in optimal cutting temperature compound (OCT, Sakura #4532), and frozen in isopentane (2-methylbutane, Sigma, 270342), followed by storage in liquid nitrogen for further use. The entire procedure was performed within 10 min. The fresh snap-frozen dissection aortic tissue was cryosectioned cut vertically (10 μm) using a Leica CM1950 cryostat (Leica, 14047742459) at −10°C. Typically, we should ensure the reproducibility of the same type and good quality of tissue morphology. RIN should be ≥7 and RNA quality assessment should be done before placing the tissue sections on visium spatial slides (**Supplementary Figure 1**).

Preparation of Quality Control Slide

The reagent kits included visium spatial tissue optimization slides and visium spatial gene expression slides, which were used for tissue optimization and spatial gene expression, respectively. For quality control experiments, poly-T20VN oligonucleotides (IDT) were uniformly spread onto Code link-activated microscopic glass slides according to the manufacturer's instructions (Stahl et al., 2016; Giacomello et al., 2017). Visium spatial tissue optimization slides contained eight capture mRNA areas with oligonucleotides, and each capture area was defined by an etched frame. Each probe had poly (dT) primers to allow the production of cDNA from polyadenylated mRNA. These probes did not contain a spatial barcode. The visium spatial gene expression slide had four capture areas (6.5 × 6.5 mm), each defined by a fiducial frame (fiducial frame + capture area is 8 × 8 mm). Every capture area contained ∼5,000 gene expression spots of RT-primers with unique barcode sequences. Each spot had a diameter of 50 μm (corresponding to a tissue domain). The center-to-center distance was 100 μm.

Surface primer for spatial arrays:

5′-CTACACGACGCTCTTCCGATCT-NNNNNNNNNNNNNNNN-NNNNNNNNNNNN-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′

Tissue Optimization (TO)

Fixation, Staining, and Imaging

The transported slides were placed on dry ice at −80°C and placed on a slide with tissue on a pre-warmed 37°C thermocycler adapter (10x genomics, 3000380). Then, the fixed tissues were fixed using ice-cold 100% methanol (Sigma, 34860) for 30 min and were stained with hematoxylin (Agilent, S330930-2) and eosin Y (Sigma, HT110216) (H&E) diluted 1:9 in 0.45M pH 6.0 Tris-buffer (Fisher, BP152-500) for 7 min and 1 min at room temperature, respectively. Between H&E staining, the glass slides were briefly dried, and bluing buffer (Agilent, CS70230-2) was added and washed off using RNase – and DNase – free Milli-Q water for 2 min. Then, we incubated the slide on the thermocycler adaptor with the thermal cycler (Thermo Fisher Scientific, 4375786) lid open for 5 min at 37°C and proceed to bright field imaging using a Leica SCN 400 slide scanner.

Tissue Permeabilization and Fluorescent cDNA Synthesis

The TO slides were placed in the slide cassette (10x genomics, 3000433) which was assembled using a slide alignment tool (10x genomics, 3000433) and was incubated with 70 μL permeabilization enzyme (10x genomics, 2000214). The positive control well includes reference RNA without any tissue. The negative control well (D2) has a tissue section not exposed to permeabilization reagents. Permeabilization times refer to the length of time (30, 24, 18, 12, 6, 3 min) tissue sections are exposed to permeabilization reagent. After permeabilization, we prepared a fluorescent reverse transcription (RT) master mix (nuclease-free water, Ambion, AM9937; RT reagent C, 10x genomics, 2000215; template switch oligo, 10x genomics, 3000228; reducing agent B, 10x genomics, 2000087; RT enzyme D, 10x genomics, 20000216) on ice according to manufacturer's instructions, and placed a thermocycler adaptor in the thermal cycler 45 min for fluorescent cDNA synthesis.

The template switch oligo was as follows:

5′-AAGCAGTGGTATCAACGCAGAGTACATrGrGrG-3′

Tissue Removal and Slide Imaging

Tissue removal was performed using a tissue removal mix (tissue removal buffer, 10x genomics, 2000221, tissue removal enzyme, 10x genomics, 3000387) which was incubated in the thermal cycler based on protocol. Then, we removed the slide from the slide cassette and centrifuged it for 30 s in a slide spinner. Fluorescence imaging was performed to all captures areas together under the same fluorescence settings using a Leica DMi8 fluorescence microscope.

Visium Spatial Gene Expression

Fixation, Staining, Imaging, Permeabilization, and RT

The sections were fixed, stained, and bright field imaging was performed as described previously; the process of tissue removal was skipped. Next, we added the permeabilization enzyme on top of the tissue, which was determined by the optimization conditions. RT mixtures used for spatial arrays (intended for library preparation and sequencing) were different from the RT mixtures used for the optimization of spatial arrays (intended for library preparation and sequencing). Then, the second strand mix was prepared (second strand reagent, 10x genomics, 2000219; second strand primer, 10x genomics, 2000217; second Strand enzyme, 10x genomics, 2000218) on ice and added to the slide incubated on the thermal cycler. Subsequently, 0.08M KOH (Sigma, 1002868722) was added to denature the second

strand, which was then collected. To determine the number of PCR cycles required for indexing, 1 μL of the purified cDNA was mixed with 9 μL of qPCR mixtures (KAPA SYBR FAST qPCR master mix, KAPA Biosystems, KK4600; cDNA primers, 10x genomics, 2000089). Then, qPCR amplifications were performed using a qPCR instrument (Applied Biosystems, 4471087), followed by cDNA amplification and quality control, purification, and transfer to separate tubes. After capturing and reverse-transcribing mRNA, we constructed a spatial gene expression library.

The second strand primer was as follows:
5′-AAGCAGTGGTATCAACGCAGAG-3′
cDNA primers:
Forward primer: 5′-CTACACGACGCTCTTCCGATCT-3′
Reverse Primer: 5′-AAGCAGTGGTATCAACGCAGAG-3′

Spatial Library Construction and Sequencing

A spatial library was prepared with 10x genomics following the user guide provided. First, fragmentation, end repair, and A-tailing were performed. The obtained cDNA profile could vary; thus, a fragmentation mix had to be prepared on ice. The amplified-cDNA was then fragmented, ligated with the adapter and sample index, and selected using SPRI beads (Beckman Coulter, B23318) to an average size of 300 bp. The quality of the libraries was evaluated at two points during the process: first, by analyzing the fragment lengths and library concentration after ligation cleanup, and second, by analyzing the library amplifiability after the final cDNA synthesis on an Agilent bioanalyzer high-sensitivity chip (Jemt et al., 2016). The constructed library was sequenced on an Illumina Nova 6000 platform.

Processing and Mapping of ST Raw Reads

Paired end 150 bp sequencing was performed using Illumina's NOVA 6000 platform. The library was sequenced using paired-end 150 bp paired-end reads using Illumina's Nova 6000 platform. Following demultiplexing the Illumina sequencer's base call files (BCLs) for each flowcell directory and converted BCLs files to FASTQ files using Bcl2Fastq2 Conversion Software (v2.20). Subsequently, the converted FASTQ file was subjected to quality control, and low-quality reads (including reads with higher N content) were filtered out. Then, the read 1 and read 2 FASTQ files were trimmed using Cutadapt (version 1.16). The read 1FASTQ file was trimmed to only the linker sequence with a length of 28 bp, and the read 2FASTQ file was only 120 bp in length; the rest files were deleted because they did not require subsequent analysis. To generate spatial feature counts for a single library using automatic fiducial alignment and tissue detection, the trimmed reads were processed with the Space Ranger pipeline (version 1.0.0) with the following arguments: "–sample V19N13_040_A1_20200728NC –slide V19N13-040 –area A1 –localcores 20 –localmem 64 –image mmexport1596289888734.jpg."

To compare the results of automatic alignment and manual alignment, a tissue assignment json file was generated in Loupe Browser, and Space Ranger count was run with "–sample V19N13_040_A1_20200728NC –slide V19N13-040 –area A1 –localcores 20 –localmem 64 –image mmexport1596289888734.jpg –loupe-alignment V19N13-040-A1.json" arguments. The reference genome used in the two Space Ranger runs was the GRCh38 v93 genome.

Selection of Methods for ST Data Analysis

The gene-spot matrices generated after ST data processing and Visium samples were analyzed using the Seurat package (version 3.1.3) in R (Butler et al., 2018). To explore the differences in normalization methods, SCTransform and log normalization were performed separately, another covariate was used to calculate the correlation of features that were grouped into groups using the Group Correlation function (settings: min.cells = 5, ngroups = 6), and the correlation between their results and the number of UMIs was tested. The results obtained by the normalization method with better correlation were selected for PCA and ICA. Then, the first 20, 30, and 50 elements analyzed by PCA and ICA were selected for subsequent analysis. Clustering of each spot is based on K-Nearest Neighbor algorithm. The distance from each point to other points was calculated first, and the shared nearest neighbor (SNN) graph was constructed according to the distance between sample points. Finally, the FindClusters function was used to determine the cluster (FindNeighbors settings: reduction = "pca/ica," nn.method = "rann," dims = 1:20/30/50, k.param = 20; Find Clusters settings: resolution = 0.8, method = "matrix," algorithm = 1). For clustering and re-dimension-reduction through uniform manifold approximation and projection (UMAP) (Becht et al., 2018) and t-distributed stochastic neighbor embedding (t-SNE) methods (van der Maaten and Geoffrey, 2008). The two methods of dimensionality reduction were evaluated based on the clustering of spot types.

Identification of Cluster-Specific Genes

For each cluster that was identified, the differentially expressed genes (DEGs) were determined in relation to all other spots. A spatial cluster gene list was first generated for all genes differentially expressed in ST clusters (average logFC $>$ 0.25, adjusted p-value $<$ 0.05, and only return positive genes). The mean expression of each gene was calculated across all spots in the cluster to identify genes that were enriched in a specific cluster. Each gene from one cluster was compared with the average expression of the same gene from the spots of all other clusters. The genes were ranked according to their expression differences, and the DEGs with the largest changes in each cluster were checked and visualized using heat maps.

Identification of Cell Types

Two databases have been used to identify cell types at different levels. First, the CellMarker database (Zhang et al., 2019) was used as a reference to classify the cell subpopulations from the annotations of cluster-specific genes. All human cell types and

their corresponding marker genes in the CellMarker database were downloaded and integrated into a dataset. Then, the cluster-specific genes in the sample performed a hypergeometric test on the dataset with the help of the enricher function in the clusterProfiler (version 3.12.0) (Yu et al., 2012a) (settings: *p*-value Cutoff = 0.005). The different cell types annotated for each cluster were finally determined according to their enrichment factors and artificial corrections. Then, the count of gene expression on each spot was compared with the Human Cell Landscape (HCL) database using the scHCL function in order to identify the Cell types that might be contained at different locations of the sample (Han et al., 2020) (settings: numbers_plot = 10).

Gene Functional Annotation

For the DEGs identified in each cluster, cluster Profiler (version 3.12.0) (Yu et al., 2012b) was used to perform Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotation, which supports statistical analysis and visual expression of the functions of genes and gene clusters. The Cluster Profiler package provides enriched GO and enriched KEGG functions to perform enrichment tests for gene ontology terms and KEGG biological pathways based on hypergeometric distribution. To reduce the false discovery rate in multiple testing, we chose FDR-corrected *p*-value less than 0.05 as the threshold.

Multi-Color Immunofluorescence Staining

The tissue was collected and prepared for the OCT-embedded frozen tissues and 8-mm-thick serial sections were prepared for. The confirmation of cell types was analyzed using Opal 7-Colour Manual IHC Kit (PerkinElmer, United States) according to the manufacturer's protocol. In brief, antigen was retrieved by AR9 buffer (pH 6.0, PerkinElmer, United States) and boiled in the oven for 15 min. After a pre-incubation with blocking buffer at room temperature for 10 min, the sections were incubated at room temperature for 1 h with mouse anti-human CD31 (Abcam 9498, United Kingdom, 1:200), rabbit anti-human CD163 (Abcam 9519, United Kingdom, 1:1000), rabbit anti-human CALD1 (Abcam 32330, United Kingdom, 1:300), rabbit anti-human HLA-DR (Abcam 92511, United Kingdom, 1:100), rabbit anti-human ACTA2 (Abcam 124964, United Kingdom, 1:300), and mouse anti-human ELN (Abcam 9519, United Kingdom, 1:100). A secondary horseradish peroxidase-conjugated antibody (PerkinElmer, United States) were added and incubated at room temperature for 10 min. Signal amplification was performed using TSA working solution diluted at 1:100 in 1 × amplification diluent (PerkinElmer, United States) and incubated at room temperature for 10 min. The other validations by multi-color IHC were performed using the same protocols with different primary antibodies as follows. The multispectral imaging was collected by Mantra Quantitative Pathology Workstation (PerkinElmer, CLS140089) at 20 × magnification and analyzed by In Form Advanced Image Analysis Software (PerkinElmer) version 2.3. For each section, a total of 5–10 high-power fields were taken based on their tissue sizes.

RESULTS

Patient and Tissue Spatial Gene Expression Information

We randomly selected one AAD patient (involving ascending aortic) with hypertension (male, 50 years old) from the first affiliated hospital of Xinjiang Medical University, who was well characterized and had a typical phenotype of AAD based on computed tomography angiography (CTA) results. The demographic data, operative details, and microarray data of the tissue are presented in **Supplementary Table 1**. **Supplementary Table 1** summarizes the data of the patient. Overall, 19,879 genes within 1,873 spot regions were analyzed in one tissue section, yielding a mean of 181,097 reads/spot with median gene and median unique molecular identifier (UMI) counts of 2514, respectively. The number of cells located within the tissue domain (each spot with a diameter of 50 μm) is estimated range from 3–10 for aortic section depending on if cells are longitudinal- or cross-sectioned. Longitudinally oriented aortic cells can potentially cover more than one single feature and numerous features contain different cell types such as ECs, SMCs, and fibroblasts. A schematic diagram of the experimental design and data analysis is shown in **Figure 1**. We compared different algorithms for the steps of normalization, component analysis, and dimensionality reduction analysis. According to the final cell annotation results, we evaluated the algorithm combination and created a pipeline suitable for analysis of human aortic tissues.

Quality Control and ST Sequencing Data Analysis in Aortic Dissection Tissue

We chose three indicators—the count of RNA, count of genes, and percentage of mitochondrial genes (nCount_RNA, nFeature_RNA, and percent mitochondrial)—to demonstrate the reliability of data; the spatial UMIs and gene distribution are shown in **Figures 2A–H**. Among them, some cells with >10% mitochondrial reads were filtered, and dead cells were removed (Ji et al., 2020; **Figures 2G,H**). The correlation between UMIs and genes obtained by two different normalization methods—SCTransform normalization and log normalization. The box plot of genes was divided into six groups according to their average expression levels (**Figures 2I,J**), which indicated that the SCTransform normalization method was better for fully normalizing highly expressed genes. Similarly, we applied two widely employed component analysis methods—principal component analysis (PCA) and independent component analysis (ICA)—for the first dimensionality reduction analysis. The first principal component heat maps of the top 30 genes obtained by PCA and ICA are shown in **Figures 2K,L**. We observed that only 12 genes were found to be co-expressed by the first principal component of the two methods, indicating that PCA and ICA provided significantly different results. The merits and demerits of two component analysis methods cannot be assessed at the genetic level alone. Therefore, the results of the two component analysis methods were selected for t-SNE and UMAP dimensionality reduction analysis. We also compared the effect of dimension selection of component analysis on the

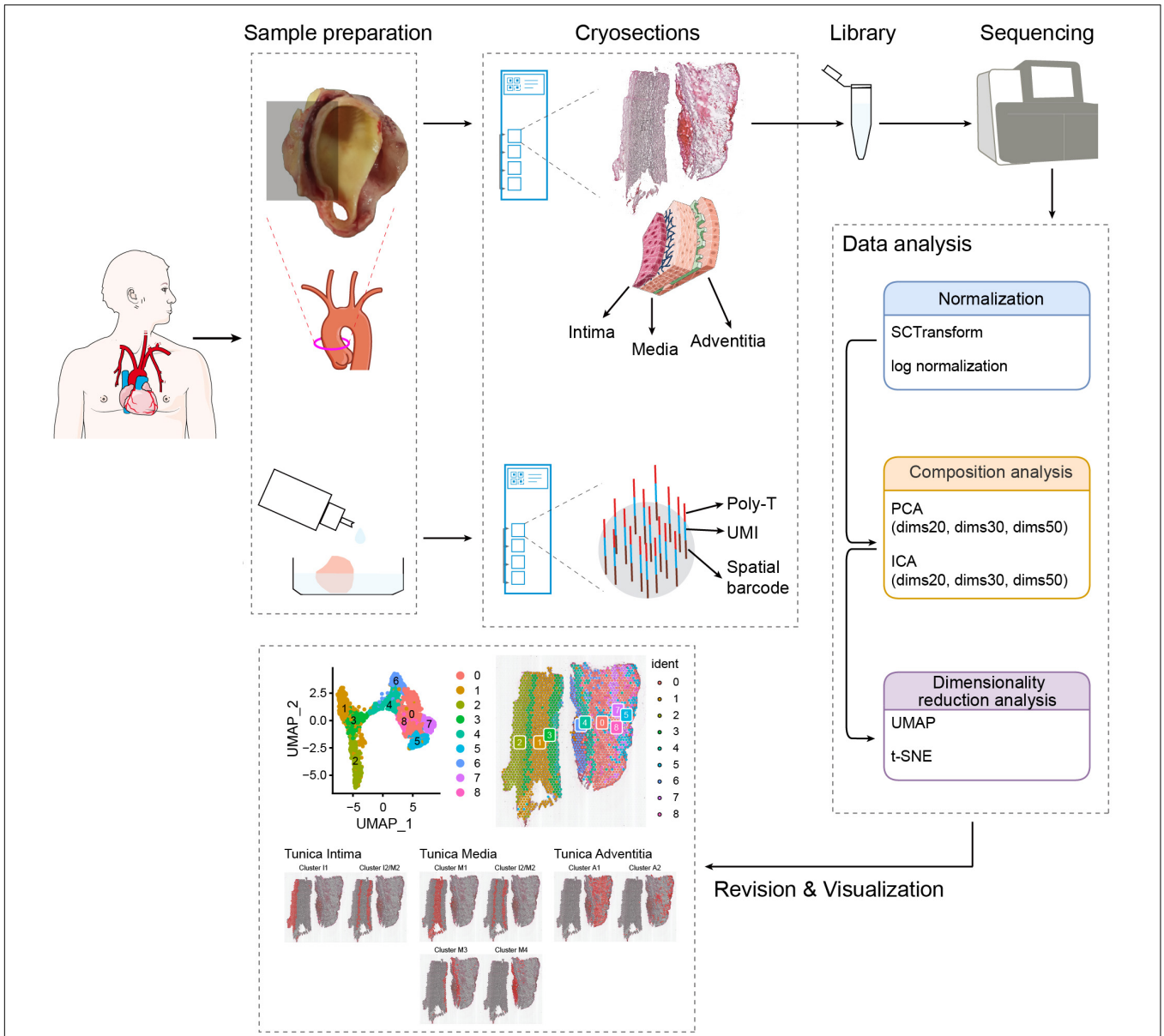

FIGURE 1 | Study design for ST in aortic dissection. Tissue of aortic dissection from the patient was dissociated and embedded in OCT; a fresh thin tissue section was obtained from cryosection, which was attached and fixed to each spot on the microarray slide; the library was sequenced and further processed to map the expressed genes to the spatial locations at which they were expressed; establishment of data analysis methods and visualization of cell populations.

results by selecting the top 20, 30, and 50 components. Finally, the optimal combination of the component analysis method and the dimensionality reduction algorithm was determined according to annotation of cell type. ICA results are shown in **Figure 2M**, and the first 20, 30, and 50 principal components were selected for t-SNE dimensionality reduction processing. The first row is the t-SNE clustering result, and the second row provides a visualization of the corresponding position of the clustering result on the tissue. Dims 20, 30, and 50 were divided into 10, 9, and 8 clusters after dimensionality reduction, respectively. The smaller the number of components selected, the more clusters were obtained on the spots. The results of t-SNE dimensionality reduction analysis of PCA are shown in **Figure 2N**. ICA and PCA provided the same number of clusters when the number of dims was confirmed. However, we found that PCA exhibited less overlap and better cluster independence than ICA in the visualization results. Similarly, we also selected different numbers of dims for UMAP clustering analysis of the two component analysis methods. The results of dims 30 are shown in **Figures 2O,P**. The UMAP dimensionality reduction methods of dims 20 and dims 50 are shown in **Supplementary Figure 2**. The visualization results of t-SNE may exaggerate the differences between cell populations and ignore the potential associations between these cell populations (**Figures 2N,P**).

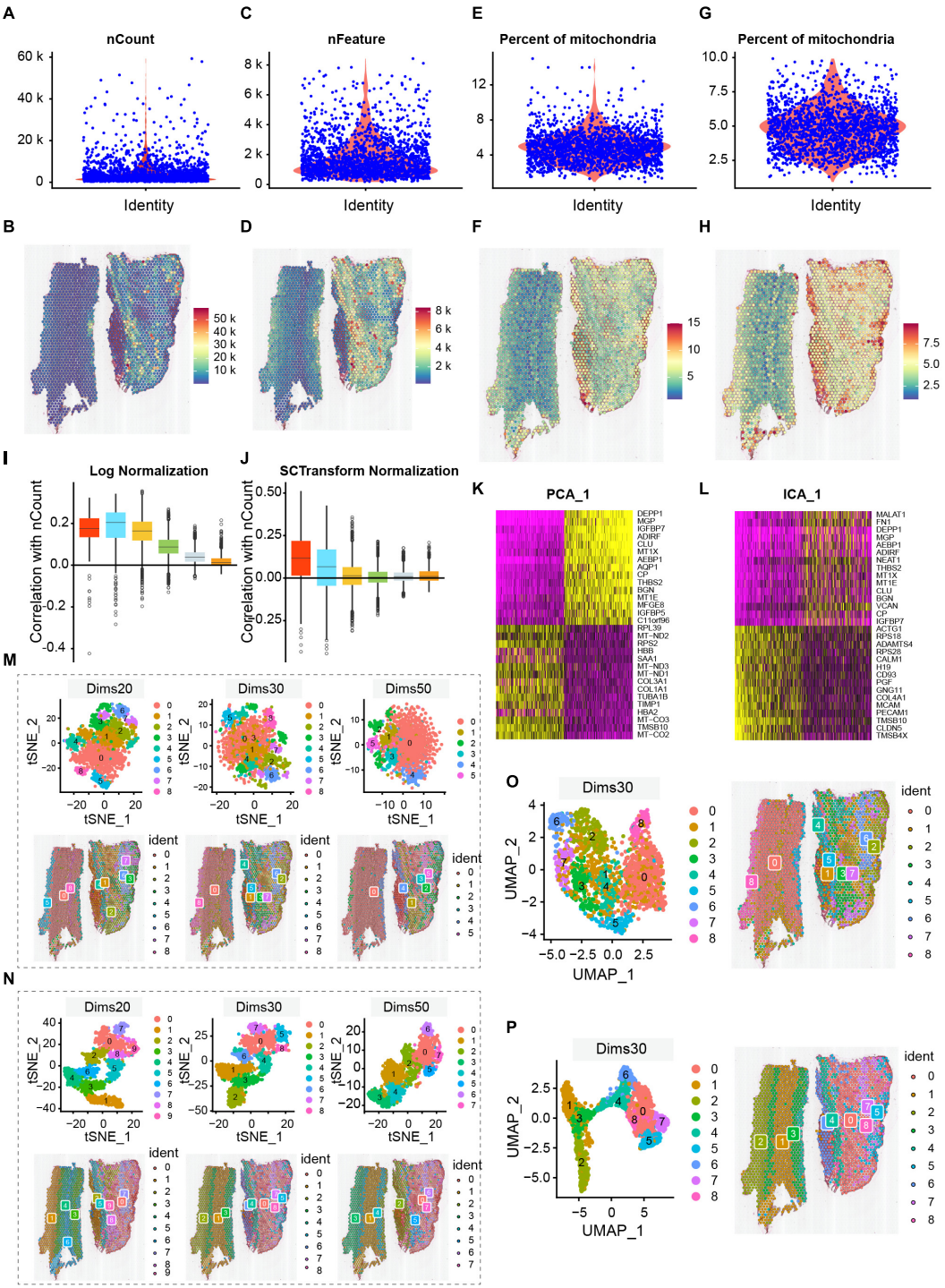

FIGURE 2 | Quality control and ST data analysis. **(A,B)** The number of nCount_RNA is range of 10,000–20,000, with the maximum not exceeding 60,000, and spatial UMIs distribution is concentrated in the aortic of tunica media and external. **(C,D)** The number of genes is mostly between 1,000 and 7,500. Combined with the distribution of UMIs data, the region with a high number of genes also had a high number of UMIs. **(E,F)** The percentage of mitochondria is low, between 1 and 12%. Correspondingly, the distribution of spatial UMIs in the tunica media and external is also rare. **(G,H)** Cells with >10% mitochondrial reads are filtered, and display distribution of spatial UMIs. The colors from blue to red represented increasing number of expression. **(I,J)** Comparison of normalization methods (log and SCTransform normalization), the SCTransform normalization is superior to Log Normalization. **(K,L)** Comparison of compositional analysis (PCA and ICA). **(M–P)** Comparison of dimensionality reduction and clustering methods, among them, **(M,O)** are under the ICA condition, the distribution of t-SNE (dims 20, 30, 50) and UMAP (dims 30); **(N,P)** are under the PCA condition, the distribution of t-SNE (dims 20, 30, 50) and UMAP (dims 30). Overall, PCA dims 30 combined with UMAP dimensionality reduction cluster analysis is an appropriate method. The Clusters are labeled using different colors.

Subsequent cell type annotation results also showed that the effect obtained by PCA dims 30 combined with UMAP dimensionality reduction cluster analysis was more consistent with the actual cell type distribution. The high quality of data guaranteed cell- and gene-level downstream analysis.

Aortic Tissue ST Sequencing Identifies Spatial Locations of Genes in the Human Aorta

The PCA combined dimensional reduction method of UMAP was performed and the first 30 principal components were annotated to obtain nine clusters in three layers of the aortic sections. We analyzed the intersection of significant genes among different clusters and found that the number of intersection genes were relatively small (**Figure 3A**). The number of genes expression in each cluster is shown in **Figure 3B**. The results showed that cluster 8 had the highest number of significant genes, followed by cluster 2; clusters 4 and 6 had a similar counts of significant genes, and cluster 1 had the lowest number of significant gene (**Figure 3B**). The top five DEG of these clusters, according to the detected mRNA transcript amounts, are displayed in a heat map (**Figure 3C**). Genes with significantly different expression were used for cell annotation. Then, we displayed the top 20 genes that were spatially related in AAD aortic tissue, as shown in **Figure 3D**. Among these spatially related genes, six genes (MGP, THBS2, AQP1, ADH1B, CFH, and CD74) were highest expressed in the intima, which are associated with the regulation of human ECs calcification and inflammation. DEPP1, IGFBP7, ADIRF, CLU, MT1X, and AEBP1 were highly expressed in both the tunica intima and media, and the proteins encoded by these genes may play a significant role in smooth muscle cell differentiation, migration, and apoptosis. Four genes (MT-CO2, MT-CO3, IGFBP5, and IGFBP3) were highly expressed in both the tunica media and adventitia, which have role in stem cell differentiation. Three genes (IGKV1D-13, SAA1, and BGN) were highly expressed in the adventitia, in which high levels of the proteins are associated with inflammatory diseases. One gene (MT-ND3) was highly expressed in the three layers of the ascending aorta; this gene may act as a transcriptional regulator for numerous genes, including some genes involved in cell metastasis and migration, and are involved in cell cycle regulation.

Distribution of Cell Types at Different Location

In order to determine cell types, we combined CellMarker and HCL database to annotate the data and compare the obtained cell types. All human cell types and marker genes in the CellMarker database were downloaded as the basis for cell identification. Cell types identified by CellMarker database are shown in **Supplementary Figure 3**; the genes for each cluster are listed in the **Supplementary Table 2**, and the position of cluster is displayed in **Figure 4A**. We found that cluster 0 and cluster 7 cell types were very similar, and cluster 5 and cluster 8 cell types were very similar. Then, the 9 clusters were merged into 7 clusters and renamed them as cluster intima (clusters I1 and I2/M2), cluster media (clusters I2/M2, M1, M3, and M4), and cluster adventitia (clusters A1 and A2) by manual annotation. Subsequently, the gene expression count on each spot was compared with the HCL database and the distribution and score of possible cell types on the spot were obtained. The major cell types (ECs, SMCs, fibroblast, and immune cells) are shown in **Figure 4B**. We found ECs located in cluster I2/M2 on the tunica intima layer, SMCs, fibroblast and macrophage located in the tunica media layer (clusters M1, M3, and M4), and ECs, fibroblast located in clusters A1 and A2 on the tunica adventitia layer. According to the HCL database, we identified the cell types and calculated its number, and found the ECs, SMCs, fibroblast, and immune cells account for a large proportion (**Figure 4C**), which was consistent with the CellMarker database. The accuracy of cell type identification was further confirmed by multi-color immunofluorescence (**Figure 4D**). The function of cell types in AAD can be inferred: cluster I1 and cluster I2/M2 cells both displayed a high correlation with differentiation, regeneration, and nerve conduction functions, such as progenitor cells, astrocytes, and microglial cells, and so on. Cells from cluster I2/M2, M1, M3, and M4 showed a high level of support and immune function, which were located in the tear position. Clusters M1 and M2 also contained numerous fibrous cells and Leydig cells, which maintain the structure of the aorta. We identified numerous types of stem cells and progenitor cells, which were closely related to vascular remodeling in clusters I1 and A1.

GO and KEGG Analysis of DEGs in Spatial Expression

After confirming the cell types in the pathological state, we next applied bioinformatics tools to determine the biological pathways affected by type A aortic dissection. We used GO gene annotation to identify cellular components and biological signals that were correlated with the spatial location in each cluster (**Figure 5A**). Pathway enrichment analysis was performed using KEGG to annotate the function (**Figure 5B**). The GO and KEGG function annotation results indicated that different types of cells at different locations showed different functional enrichment of signal pathways. For instance, cluster I1 and cluster I2/M2 were distributed in the tunica intima, which were involved in the regulation of oxygen levels and cellular activity in biological process. KEGG analysis enriched several pathways involved in the oxygen regulation of cell cycle activities and immunity moderation in the two clusters. We observed that the genes in cluster M1 and cluster I2/M2 in tunica media regulated muscle contraction and extracellular matrix activity. Similarly, pathways (vascular smooth muscle contraction, ECM-receptor interaction) enriched by KEGG were related to a highly specialized cell whose principal function was contraction. Cells in cluster M3 and cluster M4 were located in the tunica media of the tear, and involved numerous immune cells such as neutrophils, and T cells and so on in biological process. Interestingly, KEGG also enriched some pathways related to antigen processing and presentation, apoptosis, and focal adhesion, which play essential roles in

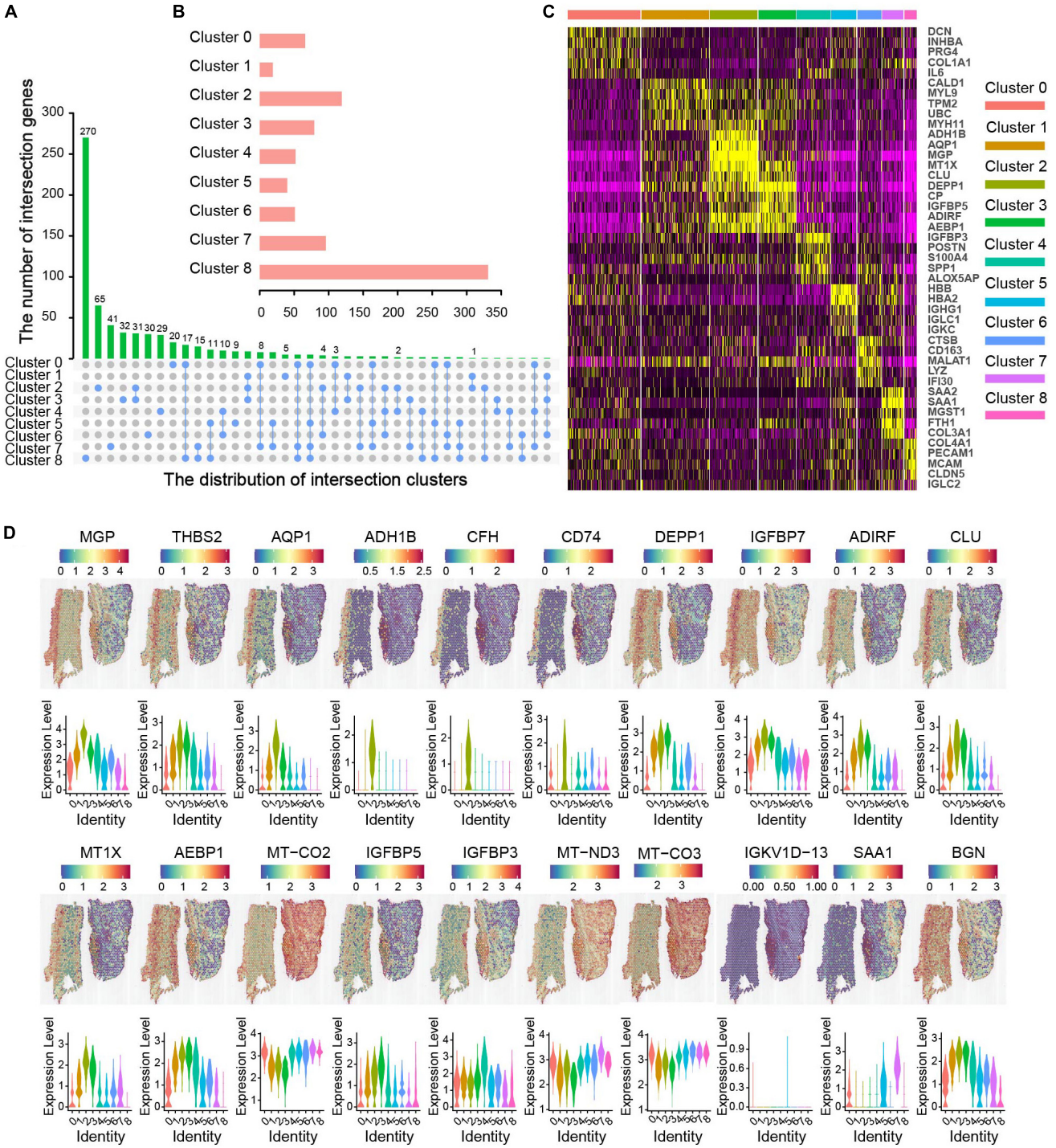

FIGURE 3 | Molecular characteristics related to spatial location in ascending aortic dissection tissue. **(A)** Correlation clusters of overlapping DEGs. The column length represents the number of overlapping genes. **(B)** The bar chart represents the number of genes in each cluster. **(C)** The heat map displays the top five (by average log [fold change]) genes in each cluster. The *X*-axis represents the distribution of intersection clusters; *Y*-axis represents the number of intersection genes. **(D)** The top 20 genes of location and expression level were in each layer. Violin plots of gene expression levels show different clusters with different colors.

cell motility, cell proliferation, cell differentiation, regulation of gene expression, and cell survival. GO analysis showed that the genes involved in neuropathic diseases and vascular functions were present in clusters A1 and A2. Likewise, genes involved in several neuroregulatory pathways that contribute to neuron degeneration, mitochondrial dysfunction, and oxidative stress

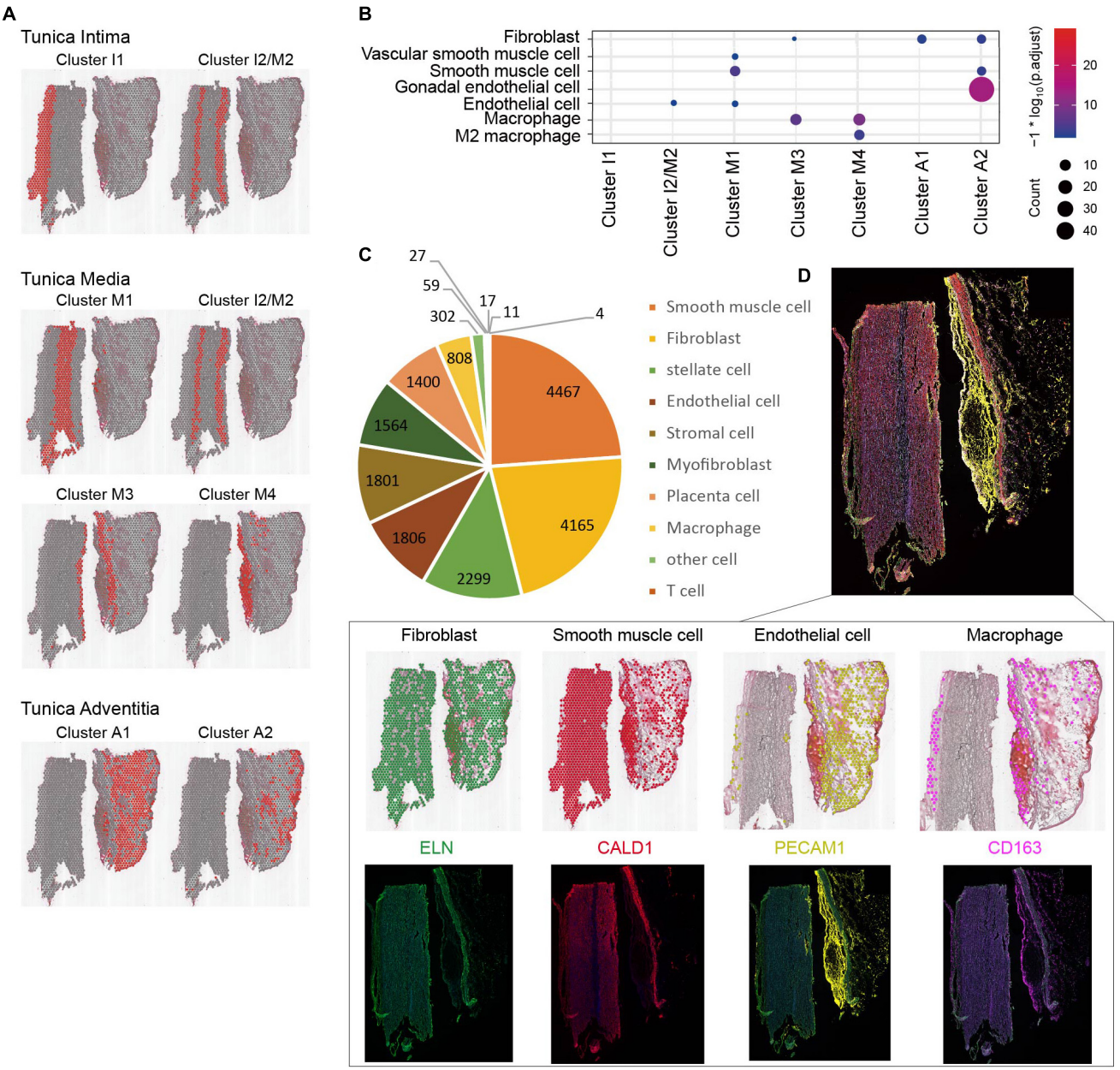

FIGURE 4 | Cell types identification of aortic dissection tissue. **(A)** The spatial location distribution of cell types in each cluster. Red region represents each cluster. **(B)** Major cell types by CellMarker database. **(C)** The major cell types and its number calculated by HCL database. **(D)** Verify the accuracy of cell types by multi-color immunofluorescence.

in the tunica adventitia were observed by KEGG enrichment analysis. In summary, the GO database screened for genes involved in biological processes in cells, and their functions were completely consistent with KEGG enriched pathways.

Visualized Gene Expression Patterns in Aortic Dissection Tissue

We retrieved 30 genes related to aortic dissection, and displayed the top 16 genes closely related to location information based on gene expression higher than 1.5, which are shown in **Figure 6**. The visual gene expression profile shows that, six genes were highly expressed in three layers—*TAGLN*, *ACTA2*, *CD44*, *FBN1*, *MMP2*, and *LOX*; three genes were highly expressed in the tunica intima and media—*CD68*, *MYH11*, and *MYLK*; and seven genes were highly expressed in the tunica adventitia—*ADAMTS1*, *ADAMTS4*, *CS*, *FKBP11*, *MVP*, *PTX3*, and *STAT3*. The genes closely associated with the pathogenesis of aortic dissection were also searched (Akutsu, 2019), and the aortic dissection tissue presents top 22 and top 21 genes of the location information

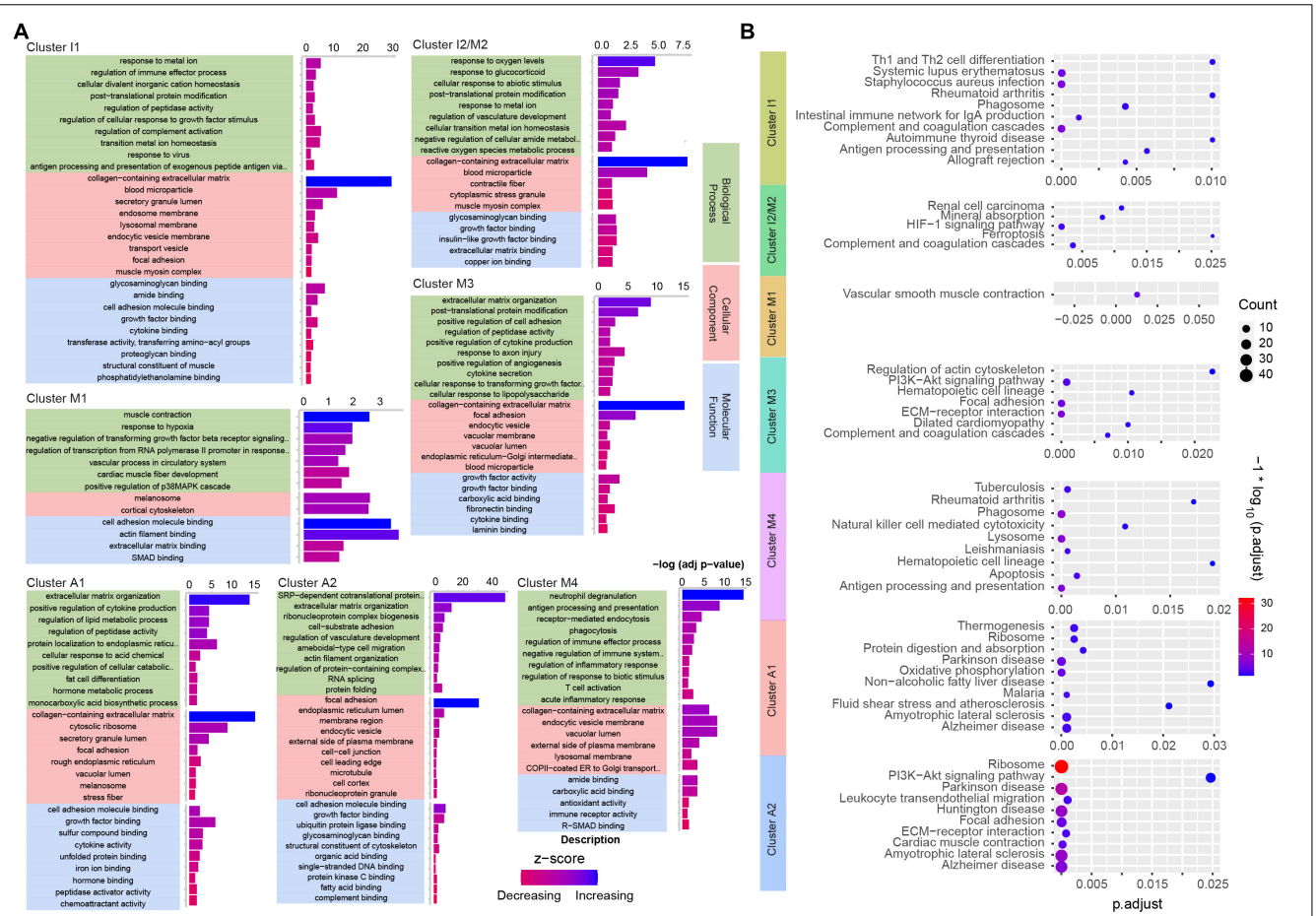

FIGURE 5 | Analysis of GO and KEGG enrichment of clusters in ascending aortic dissection tissue. **(A)** Analysis of genes enriched by GO terms (biological processes, cellular component, molecular function) (p. adjust) in each cluster. The function is reflected by *z*-score and up color-coded from red to blue. **(B)** KEGG analysis for hallmark genes of enriched pathways in each cluster. The gradient color represents the *P*-value; the size of the black spots represents the gene number.

based on the gene expression higher than 1.0 in **Figure 6**, respectively. We found that genes related to hypertension—such as *ACE*, *ANG*, *CAD*, *BMPR2*, and other genes—were also significantly expressed in ascending aortic dissection tissue reported in the literature. Among these genes, expressed in all three layers of the aorta were *PTEN*, *PKD2*, *WNK1*, *CTNNB1*, *ROCK2*, *APOE*, *SSB*, *HGS*, and *BMPR2*. The genes expressed in the tunica intima and media of the aorta were *CP*, *PVR*, and *C3*. The genes expressed in the tunica adventitia of the aorta were *FABP4*, *FH*, *ASL*, *CXCR4*, *MVD*, *APOL1*, *SCD*, *CAD*, and *ACE*. Medical history was reviewed and showed that the patient had hypertension upon admission. We also found that atherosclerotic genes were expressed in the three layers of the ascending aorta: *ABCA1*, *LDLR*, *CD40*, *APOE*, *TLR2*, *HMGB1*, *CCL2*, *UBA1*, and *ADAMTS4*; genes that were highly expressed in the tunica intima and medial were *ALDH2*, *CD47*, *TIMP3*, and *NEXN*, and genes that were highly expressed in the tunica adventitia were *CAD*, *CD36*, *TLR4*, *FH*, *RBP4*, *ACE*, *ADAM10*, and *VLDLR*. Combined with the patient's history, we found that hypertension did cause aortic dissection, and atherosclerosis was an important risk factor for aortic dissection, a result consistent with those of previous studies. However, genes associated with diabetes, inflammation, oxidative stress, and dyslipidemia were less expressed in the aortic dissection. The genes position information of these disease-causing factors in the aortic dissection tissue is listed in the **Supplementary Figure 4**.

DISCUSSION

AAD is a severe vascular disease with high mortality and limited therapeutic options (Nienaber and Clough, 2015). Understanding the biological functions, networks, and interactions of the different cell types that regulate aortic and AAD development requires both cellular information and a spatial context (Asp et al., 2019). Consequently, a visium spatial gene expression solution has been proposed to the study human aortic dissection. Here, for the first time, we provided a pipeline for aortic tissue separation and data quality control of aortic cell types through ST and showed that the pipeline can be applied to human fibrous aortic tissues with low RNA expression levels in different layers. We also preliminarily depicted a molecular

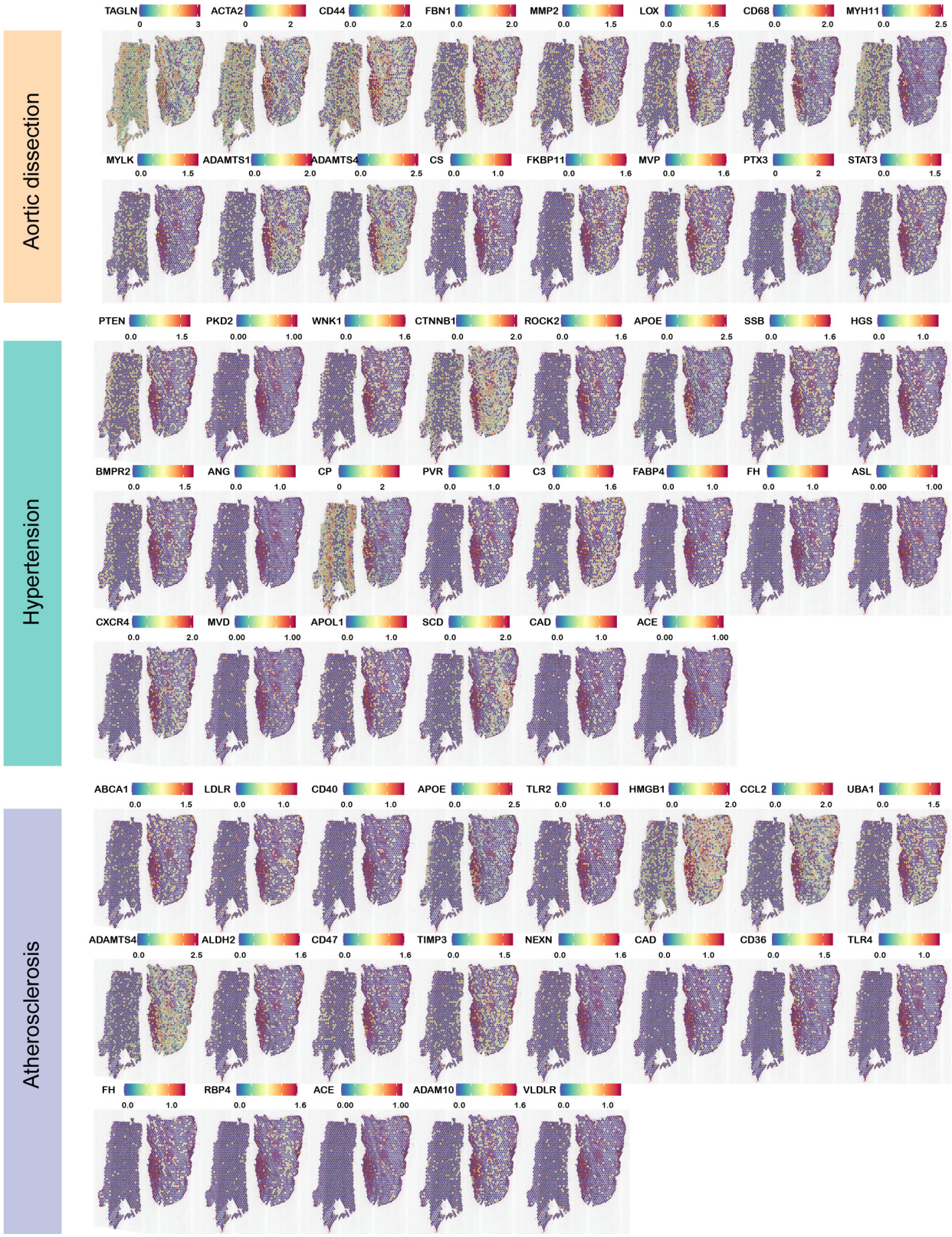

FIGURE 6 | Expression of highly expressed genes in dissection-related pathogenic factors in aortic dissection. Aortic dissection has reported the expression of highly expressed genes (hypertension, atherosclerosis) in dissection tissues of pathogenic factors. ST profiles of hypertension and atherosclerosis are listed.

landscape for ascending aortic dissection of the three layers of the aorta. Furthermore, we displayed the positional information of genes related to pathogenic factors in the aortic tissue and elaborated on the expression patterns of signal pathways in different aortic cell clusters.

The most challenging issue with the visium spatial gene expression solution is the total RNA extraction and fluorescence capture of the ascending aorta tissue (compared with other human tissue types) because the vascular tissue contains a low density of cells and a large proportion of fibrous tissue, due to which performing experiments becomes difficult. Rigorous precautions must be taken to avoid degradation of RNA during its dissociation, thereby impairing both RNA quality and yield. A further complication is that in standard RNA-seq, whole tissue biopsies are homogenized and average representations of expression profiles within the entire sample are obtained. Consequently, information on spatial patterns of gene expression is lost and signals from subpopulations of cells with deviant profiles, such as those with low-level gene expression in the tear and dysfunctional tunica medial, are obscured. To overcome these deficiencies, we aimed to analyze the gene expression in different layers of human AAD tissues using a novel ST sequencing, which allows more refined analysis of gene expression in a tissue section. We covered 1,873 spot, detected 19,879 genes, and simultaneously associated gene expression with specific cell types.

Besides, different computational methods were compared in our analysis process to identify the best processing pipeline. To our knowledge, because of the high dimensionality of ST data, differences in gene length and genome coverage, and experimental errors in processes such as cell lysis and RT, the standardization of preliminary data is critical to the interpretation of subsequent analysis results. **Figures 2I,J** show the correlation between each gene and the number of UMIs. We grouped the genes according to their mean expression and boxplots of these correlations. Log-normalization failed to adequately normalize the genes in the first four groups, which indicates that technical factors continue to affect the normalized expression estimates of highly expressed genes. On the contrary, SCTransform normalization substantially reduces this effect. The normalized data were analyzed by PCA and ICA. PCA assumes that the original components are unrelated to one another and orthogonal, and ICA assumes that the original components are independent of one another. Both were used to identify the cell types contained in the populations. To avoid overcrowding among clusters and obtain the optimal cell clustering, PCA and ICA dimension-reduction should be performed by clustering and re-dimension-reduction analysis using t-SNE and UMAP algorithms. **Figures 2N,P** show that the UMAP algorithm retains more global structures than t-SNE, especially the continuity between cell subsets. The specific pipeline enables the aortic histomorphology to map the corresponding spatial location more effectively according to the results of cell annotation.

The artery includes an abundance of multifunctional cell populations, with each of them distinctly involved in cardiovascular diseases, such as atherosclerosis and aortic dissection. Visium spatial gene expression solutions cannot reach the resolution of a single cell, which is an inevitable technical problem. We can only classify genes according to the gene expression pattern, and then describe the cells that may be contained in each group according to the existing marker genes. Therefore, two databases were used to identify the cell types and their accuracy was verified by multi-color immunofluorescence. We first presented an appropriate approach to visualize the spatial transcriptional atlas of cell types. Hence, individual transcriptomes received from each feature will provide spatial gene expression profiles. We then identified their heterogeneity in human ascending aortic dissection, which enabled the analysis of various cells corresponding to specific genes, location distribution, and functions in the tissue section. Various studies have used scRNA-seq to delineate the heterogeneity of vascular cells, including VSMCs (Dobnikar et al., 2018), ECs (Kalluri et al., 2019), macrophages (Cochain et al., 2018), and aortic adventitia cells (Gu et al., 2019) in healthy and diseased state of arteries. The ST sequencing data were analyzed using the combined SCTransform normalization, PCA dim 30, and UMAP dimensionality reduction clustering method to annotate cell types. We provided characteristic changes in the three major vascular cell types (vascular structural correlation cells, vascular development correlation cells, and immune cells) according to distinct functions in seven clusters in the three aortic layers. Consistently, both vascular resident cells, including SMCs, fibroblasts, and ECs, and infiltrating immune cells, including macrophages, B cells, T cells, and dendritic cells, were observed (Hadi et al., 2018). In the tunica intima, we identified many granulosa cells, microglial cells, and ECs, which were different from those in healthy aorta. This is associated with inflammatory infiltrating of the arterial intima, weakened vascular walls, degradation of the cytoplasmic matrix, and endothelial cells eliciting an immune response that regulates blood flow and recruits immune cells. There is a need for complementary research in the field to further highlight and compare the results with those from other locations in aorta. In the tunica media, the cell types identified were mainly SMCs and VSMCs, which have specialized functions of maintaining a stable vascular structure. These results are similar to those reported by Zhao et al. (2020). Based on the characterized transcriptomic profile, immune cells accounted for more than 80% of the total cells in the tear of tunica media, which might have important functions in cell activation in response to shear stress of blood pressure. Consistent with the results in the heart, healthy large blood vessels appear to have more endothelial cell heterogeneity, whereas mural cells exhibit less transcriptional variability (Chavkin and Hirschi, 2020). We also identified numerous stem cells; it is also possible that adventitial stem cells or myofibroblasts may transdifferentiate into a contractile phenotype and migrate into the tunica media (Pedroza et al., 2020). Adding another layer of diversity to the cellular landscape of tunica adventitia, gonadal endothelial cells and neuronal cells were detected by ST sequencing despite the rarity of resident macrophages, which attract immune cells. The largest population of cells was of stem cells, which control and maintain cell regeneration and play an important role in angiogenesis and remodeling (Baron et al., 2018; Brown et al., 2018).

In the diseased state, these adaptive changes do not return to baseline levels but instead initiate pathological vascular alterations observed with AAD. We revealed that DEGs included those related to cellular activity (*AEBP1*, *ADIRF*, and *IGFBP5*), inflammation (*MGP*, *BGN*, and *SAA1*), and neurons (*CLU*, *MT-CO2*, and *ADH1B*), as well as the genes and feature signaling pathways for each cluster. The KEGG pathway annotation showed that all DEGs were significantly enriched in multiple pathways, including 15 in the tunica intima (clusters I1 and I2/M2), 21 in the tunica media (clusters M1, I2/M2, M3, and M4), and 20 in the tunica adventitia (clusters A1 and A2). They play an important role in the process of VSMCs loss or dysfunction, proteoglycan accumulation, and collagen and elastic fiber cross-linked disorder and fragmentation. Additionally the results were consistent with that observed for the heterogeneity of cell type functions, which serves as a direct evidence for subsequent study. Preliminary result demonstrating the pathways suggests that AAD is among the complex mechanisms through which they participate in vascular injury repair and thus is a potentially interesting field.

There were some limitations in our study. First, as this study contains a limited number of subjects no conclusions about AAD disease progression can be made. More samples are required to elaborate potential mechanism underlying the interactions between cells and aortic dissection. Second, while ST sequencing permits simultaneous characterization of cell type within the aorta, this data provides a limited view of the true functional changes in AAD pathogenesis that are undoubtedly affected by cellular processes other locations of aorta Other positions are required to enhance the reliability of the results, including aortic arch, the left common carotid and the left subclavian artery.

CONCLUSION

We provided a reliable ST sequencing data computational method available for the scientific community to further explore the key factors and pathways involved physiologically in the low cell density and high fiber of the aorta. The pipeline was applied to cell annotation and pathway enrichment analysis corresponding to cell location and our findings may provide insights into the function and regulation of AAD onset and progression and pave the way for selective targeting of causative cell populations in vascular diseases.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: PRJNA730333.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee at the First Affiliated Hospital of Xinjiang Medical University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

Y-NY and DL conceived the project. Y-NY, DL, and X-ML supervised the study. QH, ZL, Y-HL, TT, J-YL, and FL collected the samples. Y-HL and YC constructed RNA-seq libraries and carried out sequencing and designed the dynamic cross-tissue network analysis method. B-BF, DA, and QZ performed the cryosectioned. YC, Y-HL, and DL performed the bioinformatics and data analysis. Y-HL and YC drafted the manuscript. DL and Y-NY revised the manuscript. All the authors read and approved the final manuscript.

FUNDING

This work was financially supported by the National Key R&D Program of China (No. 2018YFC1312804), the National Natural Science Foundation of China (Nos. 82070368, 81770363), the Key R&D Projects in Xinjiang Uygur Autonomous Region (Nos. 2020B03002, 2020B03002-1, 2020B03002-2, 2020B03002-3), the Natural Science Foundation of Xinjiang (No. 2021D01C345), and the 13th Five-Year Plan of Xinjiang key discipline (No. 33-0104006020801#).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.698124/full#supplementary-material

Supplementary Figure 1 | The tissue contains RNA of good quality.

Supplementary Figure 2 | Comparison of compositional analysis and dimensionality reduction. **(A)** The first row shows a comparison PCA of UMAP dim 20 and dim 50. The second row shows the UMIs distribution of each cluster, and the different clusters are labeled by using different colors. **(B)** The first row shows a comparison ICA of UMAP dim 20 and dim 50. The second row shows the UMIs distribution of each cluster, and different clusters are labeled using different colors.

Supplementary Figure 3 | Cell types identified by CellMarker database. The identified cell types in each cluster, which are annotated based on existed marker gene information. The gradient color represents the *P*-value; the size of the black spots represents the gene number.

Supplementary Figure 4 | Expression of highly expressed genes in dissection-related pathogenic factors in aortic dissection. Aortic dissection has reported the expression of highly expressed genes in dissection tissues of pathogenic factors (diabetes, dyslipidemia, inflammation, oxidative stress). These factors genes of ST profiles are listed.

Supplementary Table 1 | Basic information of patient.

Supplementary Table 2 | The genes of each cluster.

REFERENCES

Akutsu, K. (2019). Etiology of aortic dissection. *Gen. Thorac. Cardiovasc. Surg.* 67, 271–276. doi: 10.1007/s11748-019-01066-x

Asp, M., Giacomello, S., Larsson, L., Wu, C., Furth, D., Qian, X., et al. (2019). A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell* 179, 1647–1660. doi: 10.1016/j.cell.2019.11.025

Asp, M., and Salmen, F. (2017). Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Sci. Rep.* 7:12941. doi: 10.1038/s41598-017-13462-5

Baron, C. S., Kester, L., Klaus, A., Boisset, J. C., Thambyrajah, R., Yvernogeau, L., et al. (2018). Single-cell transcriptomics reveal the dynamic of haematopoietic stem cell production in the aorta. *Nat. Commun.* 9:2517. doi: 10.1038/s41467-018-04893-3

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., et al. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* doi: 10.1038/nbt.4314

Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstrahle, J., et al. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* 9:2419. doi: 10.1038/s41467-018-04724-5

Brown, I. A. M., Diederich, L., Good, M. E., DeLalio, L. J., Murphy, S. A., Cortese-Krott, M. M., et al. (2018). Vascular smooth muscle remodeling in conductive and resistance arteries in hypertension. *Arterioscler. Thromb. Vasc. Biol.* 38, 1969–1985. doi: 10.1161/atvbaha.118.311229

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Chavkin, N. W., and Hirschi, K. K. (2020). Single cell analysis in vascular biology. *Front. Cardiovasc. Med.* 7:42. doi: 10.3389/fcvm.2020.00042

Cochain, C., Vafadarnejad, E., Arampatzi, P., Pelisek, J., Winkels, H., Ley, K., et al. (2018). Single-cell RNA-seq reveals the transcriptional landscape and heterogeneity of aortic macrophages in murine atherosclerosis. *Circ. Res.* 122, 1661–1674. doi: 10.1161/CIRCRESAHA.117.312509

Dobnikar, L., Taylor, A. L., Chappell, J., Oldach, P., Harman, J. L., Oerton, E., et al. (2018). Disease-relevant transcriptional signatures identified in individual smooth muscle cells from healthy mouse vessels. *Nat. Commun.* 9:4567. doi: 10.1038/s41467-018-06891-x

Erbel, R., Aboyans, V., Boileau, C., Bossone, E., Bartolomeo, R. D., Eggebrecht, H., et al. (2014). 2014 ESC guidelines on the diagnosis and treatment of aortic diseases document covering acute and chronic aortic diseases of the thoracic and abdominal aorta of the adult. The task force for the diagnosis and treatment of aortic diseases of the European Society of Cardiology (ESC). *Eur. Heart J.* 35, 2873–2926.

Gerlinger, M., Rowan, A. J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New Engl. J. Med.* 366, 883–892. doi: 10.1056/NEJMoa1113205

Giacomello, S., Salmen, F., Terebieniec, B. K., Vickovic, S., Navarro, J. F., Alexeyenko, A., et al. (2017). Spatially resolved transcriptome profiling in model plant species. *Nat. Plants* 3:17061. doi: 10.1038/nplants.2017.61

Gu, W., Ni, Z., Tan, Y. Q., Deng, J., Zhang, S. J., Lv, Z. C., et al. (2019). Adventitial cell atlas of wt (Wild Type) and ApoE (Apolipoprotein E)-deficient mice defined by single-cell RNA sequencing. *Arterioscler. Thromb. Vasc. Biol.* 39, 1055–1071. doi: 10.1161/ATVBAHA.119.312399

Guo, D. C., Grove, M. L., Prakash, S. K., Eriksson, P., Hostetler, E. M., LeMaire, S. A., et al. (2016). Genetic variants in LRP1 and ULK4 Are associated with acute aortic dissections. *Am. J. Hum. Genet.* 99, 762–769. doi: 10.1016/j.ajhg.2016.06.034

Hadi, T., Boytard, L., Silvestro, M., Alebrahim, D., Jacob, S., Feinstein, J., et al. (2018). Macrophage-derived netrin-1 promotes abdominal aortic aneurysm formation by activating MMP3 in vascular smooth muscle cells. *Nat. Commun.* 9:5022. doi: 10.1038/s41467-018-07495-1

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., et al. (2020). Construction of a human cell landscape at single-cell level. *Nature* 581, 303–309. doi: 10.1038/s41586-020-2157-4

He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å, et al. (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Engin.* 4, 827–834. doi: 10.1038/s41551-020-0578-x

Huang, X., Yue, Z., Wu, J., Chen, J., Wang, S., Wu, J., et al. (2018). MicroRNA-21 knockout exacerbates angiotensin II-induced thoracic aortic aneurysm and dissection in mice with abnormal transforming growth factor-beta-SMAD3 signaling. *Arterioscler. Thromb. Vasc. Biol.* 38, 1086–1101. doi: 10.1161/atvbaha.117.310694

Jemt, A., Salmen, F., Lundmark, A., Mollbrink, A., Fernandez Navarro, J., Stahl, P. L., et al. (2016). An automated approach to prepare tissue-derived spatially barcoded RNA-sequencing libraries. *Sci. Rep.* 6:37137. doi: 10.1038/srep37137

Ji, A. L., Aj Rubin, K., Thrane, S., Jiang, D. L., Reynolds, R. M., Meyers, M. G., et al. (2020). Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* 182, 497–514. doi: 10.1016/j.cell.2020.05.039

Kalluri, A. S., Vellarikkal, S. K., Edelman, E. R., Nguyen, L., Subramanian, A., Ellinor, P. T., et al. (2019). Single-cell analysis of the normal mouse aorta reveals functionally distinct endothelial cell populations. *Circulation* 140, 147–163. doi: 10.1161/circulationaha.118.038362

Lundmark, A., Gerasimcik, N., Bage, T., Jemt, A., Mollbrink, A., Salmen, F., et al. (2018). Gene expression profiling of periodontitis-affected gingival tissue by spatial transcriptomics. *Sci. Rep.* 8:9370. doi: 10.1038/s41598-018-27627-3

Maniatis, S., Aijo, T., Vickovic, S., Braine, C., Kang, K., Mollbrink, A., et al. (2019). Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* 364, 89–93. doi: 10.1126/science.aav9776

Nienaber, C. A., and Clough, R. E. (2015). Management of acute aortic dissection. *Lancet* 385, 800–811. doi: 10.1016/s0140-6736(14)61005-9

Oller, J., Méndez-Barbero, N., Ruiz, E. J., Villahoz, S., Renard, M., Canelas, L. I., et al. (2017). Nitric oxide mediates aortic disease in mice deficient in the metalloprotease Adamts1 and in a mouse model of Marfan syndrome. *Nat. Med.* 23, 200–212. doi: 10.1038/nm.4266

Pedroza, A. J., Tashima, Y., Shad, R., Cheng, P., Wirka, R., Churovich, S., et al. (2020). Single-cell transcriptomic profiling of vascular smooth muscle cell phenotype modulation in marfan syndrome aortic aneurysm. *Arterioscler. Thromb. Vasc. Biol.* 40, 2195–2211. doi: 10.1161/ATVBAHA.120.314670

Stahl, P. L., Salmen, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82. doi: 10.1126/science.aaf2403

Thrane, K., Eriksson, H., Maaskola, J., Hansson, J., and Lundeberg, J. (2018). Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.* 78, 5970–5979. doi: 10.1158/0008-5472.CAN-18-0747

van der Maaten, L., and Geoffrey, H. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wang, Y., Dong, C. Q., Peng, G. Y., Huang, H. Y., Yu, Y. S., Ji, Z. C., et al. (2019). MicroRNA-134-5p regulates media degeneration through inhibiting VSMC phenotypic switch and migration in thoracic aortic dissection. *Mol. Ther. Nucleic Acids* 16, 284–294. doi: 10.1016/j.omtn.2019.02.021

Yang, J., Zou, S., Liao, M., and Qu, L. (2018). Transcriptome sequencing revealed candidate genes relevant to mesenchymal stem cells' role in aortic dissection patients. *Mol. Med. Rep.* 17, 273–283. doi: 10.1093/ejcts/ezu171

Yang, K., Ren, J., Li, X., Wang, Z., Xue, L., Cui, S., et al. (2020). Prevention of aortic dissection and aneurysm via an ALDH2-mediated switch in vascular smooth

muscle cell phenotype. *Eur. Heart J.* 41, 2442–2453. doi: 10.1093/eurheartj/ehaa352

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012a). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012b). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* 16, 284–287.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 47, D721–D728. doi: 10.1093/nar/gky900

Zhao, G., Lu, H., Chang, Z., Zhao, Y., Zhu, T., Chang, L., et al. (2020). Single cell RNA sequencing reveals the cellular heterogeneity of aneurysmal infrarenal abdominal aorta. *Cardiovasc. Res.* 117, 1402–1416. doi: 10.1093/cvr/cvaa214

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Citation: Li Y-H, Cao Y, Liu F, Zhao Q, Adi D, Huo Q, Liu Z, Luo J-Y, Fang B-B, Tian T, Li X-M, Liu D and Yang Y-N (2021) Visualization and Analysis of Gene Expression in Stanford Type A Aortic Dissection Tissue Section by Spatial Transcriptomics. Front. Genet. 12:698124. doi: 10.3389/fgene.2021.698124

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 September 2021
doi: 10.3389/fgene.2021.728333

Discovering Cerebral Ischemic Stroke Associated Genes Based on Network Representation Learning

Haijie Liu[1†], Liping Hou[2†], Shanhu Xu[3], He Li[4], Xiuju Chen[5], Juan Gao[6], Ziwen Wang[7], Bo Han[1], Xiaoli Liu[3] and Shu Wan[3]*

[1] Department of Neurology, Xuanwu Hospital, Capital Medical University, Beijing, China, [2] Department of Clinical Laboratory, General Hospital of Heilongjiang Province Land Reclamation Bureau, Harbin, China, [3] Affiliated Zhejiang Hospital, Zhejiang University School of Medicine, Hangzhou, China, [4] Department of Automation, College of Information Science and Engineering, Tianjin Tianshi College, Tianjin, China, [5] Department of Neurology, Tianjin Nankai Hospital, Tianjin, China, [6] Department of Neurology, Baoding No. 1 Central Hospital, Baoding, China, [7] Graduate School of Chengde Medical College, Chengde, China

Cerebral ischemic stroke (IS) is a complex disease caused by multiple factors including vascular risk factors, genetic factors, and environment factors, which accentuates the difficulty in discovering corresponding disease-related genes. Identifying the genes associated with IS is critical for understanding the biological mechanism of IS, which would be significantly beneficial to the diagnosis and clinical treatment of cerebral IS. However, existing methods to predict IS-related genes are mainly based on the hypothesis of guilt-by-association (GBA). These methods cannot capture the global structure information of the whole protein–protein interaction (PPI) network. Inspired by the success of network representation learning (NRL) in the field of network analysis, we apply NRL to the discovery of disease-related genes and launch the framework to identify the disease-related genes of cerebral IS. The utilized framework contains three main parts: capturing the topological information of the PPI network with NRL, denoising the gene feature with the participation of a stacked autoencoder (SAE), and optimizing a support vector machine (SVM) classifier to identify IS-related genes. Superior to the existing methods on IS-related gene prediction, our framework presents more accurate results. The case study also shows that the proposed method can identify IS-related genes.

Keywords: cerebral ischemic stroke, network embedding, disease gene prediction, PPI network, network representation learning

OPEN ACCESS

Edited by:
Jiajie Peng, Northwestern Polytechnical University, China

Reviewed by:
Xiaoke Ma, Xidian University, China
Yanshuo Chu, University of Texas MD Anderson Cancer Center, United States

***Correspondence:**
Shu Wan
wanshu@zju.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *21 June 2021*
Accepted: *26 July 2021*
Published: *01 September 2021*

Citation:
Liu H, Hou L, Xu S, Li H, Chen X, Gao J, Wang Z, Han B, Liu X and Wan S (2021) Discovering Cerebral Ischemic Stroke Associated Genes Based on Network Representation Learning. Front. Genet. 12:728333. doi: 10.3389/fgene.2021.728333

INTRODUCTION

Cerebral ischemic stroke (IS) is the most common type of stroke, which results from a sudden cessation of adequate amounts of cerebral blood supply through vessels (Sacco et al., 2013). As cerebral IS appears to be a complex disorder associated with both genetic and environmental factors, it is highly demanding to demonstrate the underlying patterns of inheritance (Matarin et al., 2010). Some IS-associated genes have been detected, verified, and recorded in recent studies (Cheng et al., 2014). Nevertheless, many unknown cerebral IS-associated genes still need to be discovered. Identifying such genes will significantly contribute to a more detailed understanding of the inherent molecular mechanism of cerebral IS, and will aid the discovery of clinical biomarkers and

therapeutic targets. With the development of statistical and machine learning methods in disease-gene discovery, it is crucial to construct and implement a promising computational algorithm for the task of effectively identifying the IS-related genes.

In recent years, predicting disease-related genes has drawn much attention in relative fields and many graph-based computational methods have performed proficiency in integrating large-scale omics data and disease phenotype (Nguyen and Ho, 2012; Zemojtel et al., 2014; Kumar et al., 2018; Wang T. et al., 2020; Peng et al., 2021b). It can be surmised that the prime cost of discovering effective drug targets will be decreased with the engagement of computational algorithms. Under the hypothesis of guilt-by-association (GBA) that most of the existing methods have relied on, it is practicable to explore and even crystallize the unknown disease genes via their connections with the known disease genes (Molet et al., 2013). Based on the GBA hypothesis, disease-associated genes are closely connected or share similar topological structure in the protein–protein interaction (PPI) network. Thus, the effective application of GBA and network-based algorithms largely depends on correct calculation of the distance or similarity between candidate genes and known disease genes.

Many network-based computational methods have also been proposed in recent years (Wang et al., 2019a,b; Yang et al., 2019). For predicting disease genes, one of the initial methods is to simply count the number of disease-genes in the neighborhood of a candidate gene (Oti et al., 2006). However, the direct neighborhood counting methods fail to capture the distant disease genes, i.e., the disease-genes not directly connecting to the candidate gene will be ignored. In this regard, several methods are proposed by considering the distances among genes in a gene network. For instance, methods calculating the shortest path length (SPL) between a candidate gene and the known disease gene have been proposed to examine their biological relatedness. However, Embar et al. (2016) have proved that the average SPL of a gene set only reveals the degree distribution of the genes and their network topology. Thus, methods relying on SPL failed to demonstrate the functional coherence as supposed (Embar et al., 2016). To overcome the shortage of single topological feature in disease-gene prediction, Xu and Li (2006) proposed a method to use multiple topological features together. They integrated five types of local topological features, including degree, 1N index, 2N index, average distance to disease-genes, and positive topology coefficient, and utilized k-nearest neighbors (KNN) as the classifier to distinguish novel disease genes (Xu and Li, 2006). Although the above methods are proven useful, the predicting performance is still not good enough. This is because these methods merely consider local topological features while ignoring the global information. The involvement of global topological information is suggested as a way for obtaining a more impressive gene node presentation and downstream outcomes (Cao et al., 2014; Vuillon and Lesieur, 2015; Peng et al., 2016, 2019).

Considering the global topology information during the learning process is deemed to cause prohibitive computational cost as well as low learning accuracy (Dai et al., 2020). Thus, some studies have tried to develop cost-efficient methods to improve the learning accuracy and explore the multidimensional interactions between genes and proteins with random walk with restart (Valdeolivas et al., 2017; Peng et al., 2019, 2021c). In a recent study, inspired by the idea from random walk with restart, we initiate further application of network representation learning (NRL) that promotes the dimensional reduction of the gene representation in the network and discover the disease-related genes of cerebral IS (Peng et al., 2021a).

In this paper, we utilize the current NRL-based algorithms to predict cerebral IS disease-related genes. Our contributions are three-fold: (1) global topological features of nodes in the PPI network are learned through three cutting-edge graph embedding methods, such as DeepWalk, LINE, and Node2Vec, and their performances are evaluated; (2) the node embeddings are transformed into a low-dimensional space using the deep learning model of a stacked auto-encoder; and (3) we show the superior performance of NRL-based methods for IS gene prediction, and novel genes associated with IS were nominated.

METHODOLOGY

We apply the NRL-based workflow, as shown in **Figure 1**, to discover the disease-related genes of IS. The workflow can be concluded into three main parts: extracting features via node representation learning, reducing feature dimension through a stacked autoencoder (SAE; Larochelle et al., 2014), and classification using support vector machine (SVM; Chang and Lin, 2011). First, we utilize three NRL-based algorithms, Node2vec (Grover and Leskovec, 2016), DeepWalk (Perozzi et al., 2014), and LINE (Jian et al., 2015) to collect the high-dimensional feature representation of each gene node from PPI network and compare those structural features captured by different algorithms. In order to avoid the influence of high-dimensional noise, next, we launch a SAE model to map corresponding feature vectors into lower dimensional space. Finally, we use an SVM classifier and convert the process of predicting disease-related genes of IS into node classification problem.

Graph Embedding for the PPI Network

Based on the need for capturing the global features of topological properties from the PPI network, three classic algorithms (Node2vec, DeepWalk, and LINE) are introduced in the following part. We learn the non-linear feature vectors for genes in the PPI network and compare the performances of the above algorithms.

DeepWalk serves as the first implemented NRL algorithm and is managed to represent nodes from the PPI network as novel latent feature vectors. At the outset, it runs the classic stochastic process to generate multiple random paths with certain length and this will formulate the topological structure. Then, it can be attributed to a natural language learning process, where the generated random paths are treated as sequences, where nodes are considered as words. Next, the skip-gram neuronal network model is utilized to maximize the probability of neighbors of the nodes in the random walk sequence. In

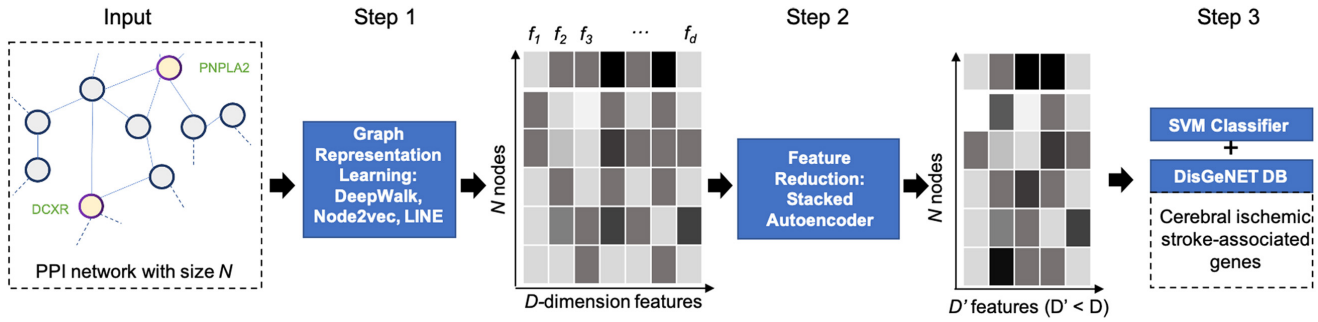

FIGURE 1 | The workflow of the proposed network representation learning (NRL) framework. The framework contains three main parts: Step 1, capturing the topological information of the protein–protein interaction (PPI) network with NRL; Step 2, denoising the gene feature using a stacked autoencoder (SAE); Step 3, training a support vector machine (SVM) classifier to predict IS-related genes.

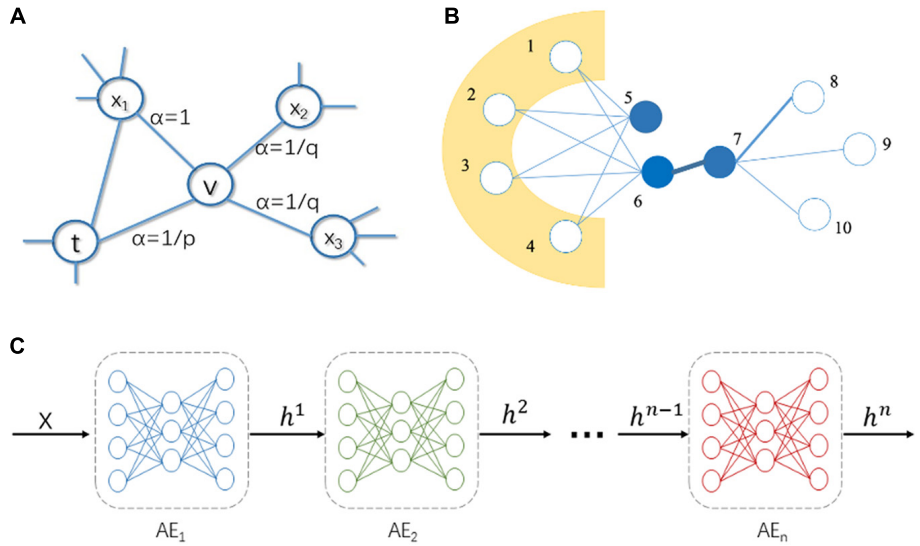

FIGURE 2 | Illustration of Node2vec, LINE, and stacked autoencoder (SAE). **(A)** The biased random walk in Node2vec. **(B)** Illustration of first-order and second order similarity in LINE. **(C)** Structure of SAE, where color represents different layer of the SAE.

the end, the weight matrix of hidden layer in the skip-gram neuronal network is used as the low-dimensional representation vectors. Node2vec improves DeepWalk algorithm by utilizing a biased random walk process to generate the random paths. It sets hyperparameters p and q to control the directions of random walk in the manner of breadth-first search (BFS) or depth-first search (DFS), thereby capturing local and global structural features in the network. The function of super parameters p and q in the random walk procedure is elucidated in **Figure 2A**. Parameter p is called the return parameter, which mainly determines the process of revisiting the nodes within random walk. When p is relatively small, the random walk is more inclined to revisit the nodes that have been visited. Parameter q is called the in-out parameter, which affects the possibility of capturing "local" or "global" nodes. When $q > 1$, the random walk is inclined to BFS, and when $q < 1$, the random walk is inclined to DFS. Intuitively, the in-out parameter q controls the ratio of performing BFS or DFS. Particularly, if p and q are both equal to 1, the Node2vec algorithm can be simply reckoned as DeepWalk.

Large-scale Information Network Embedding (LINE) is a NRL method based on the assumption of neighborhood similarity, which can be used to learn the low-dimension representation of nodes in a graph. To store network structural information, there are two different definitions of similarity between vertices in a graph. For example, in **Figure 2B**, there is a strong tie between vertex 6 and 7, so they are two similar vertices. Even if there is no direct correlation between vertex 5 and 6, they share many common neighbors (vertex 1, 2, 3, and 4), which make them the similar nodes.

The two kinds of similarity are described as first-order proximity and second-order proximity. The first-order proximity considers that the greater the edge weight of two vertices, the more similar the two vertices are. Second-order proximity considers that the more common neighbors two vertices have, the more similar the two vertices are.

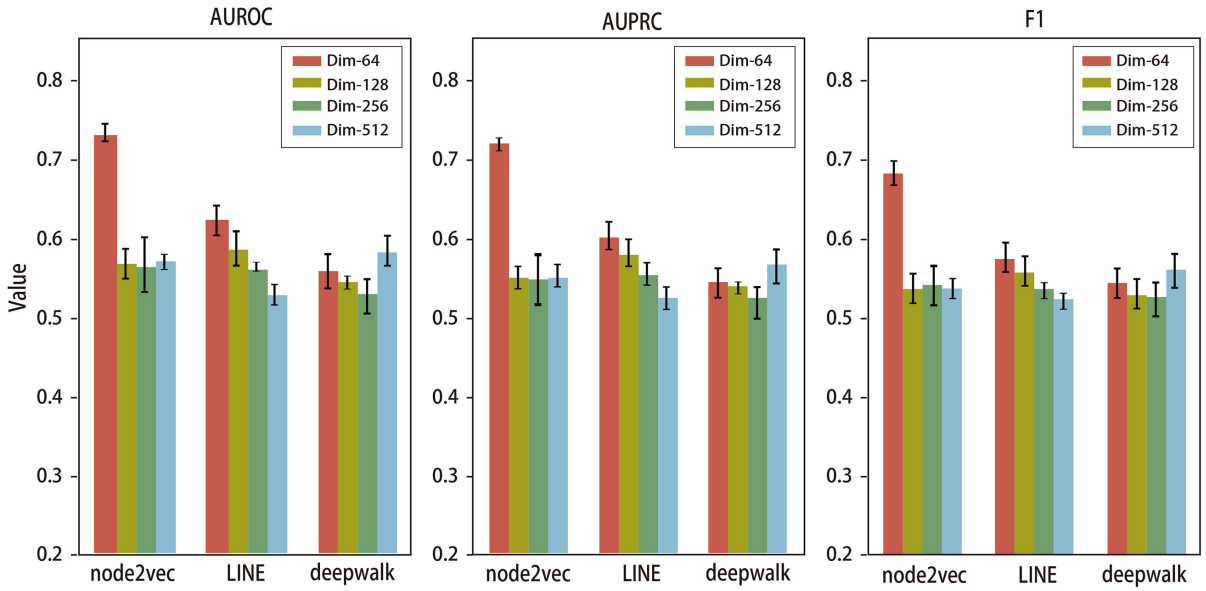

FIGURE 3 | AUROC, AUPRC, and F1 values of node2vec, LINE, and deepwalk using features in different numbers of dimensions. The error bar shows performance variation during the five-fold cross-validation.

The first-order proximity in a network is the local pairwise proximity between two vertices. The first-order proximity between u and v is equal to the weight on that edge, w_{uv}. If no edge is observed between u and v, their first-order proximity is 0. For each undirected edge (i, j), the joint probability between vertex v_i and v_j is defined as follows:

$$p_1\left(v_i, v_j\right) = \frac{1}{1 + \exp\left(-\overrightarrow{u}_i^T \cdot \overrightarrow{u}_j\right)} \tag{1}$$

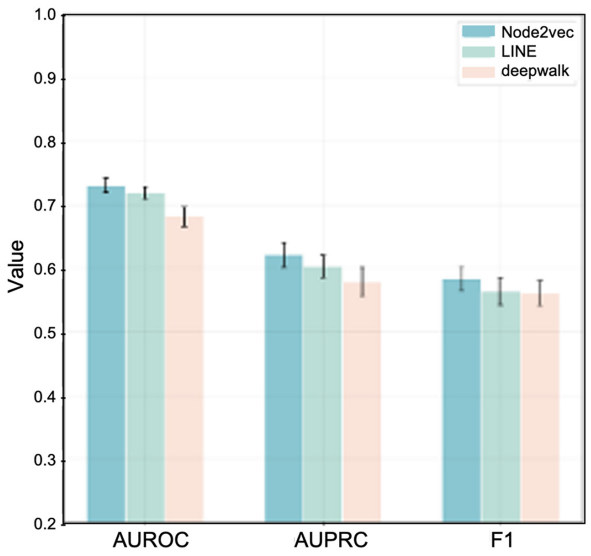

FIGURE 4 | Best performance comparison among three algorithms in the task of IS-related gene prediction.

The empirical probability is defined as $\hat{p}_1\left(i,j\right) = \frac{w_{ij}}{W}$, where W = $\sum_{(i,j)\in E} w_{ij}$. The objective function is as follows:

$$O_1 = d(\hat{p}_1\left(\cdot,\cdot\right), p_1\left(\cdot,\cdot\right)) \tag{2}$$

The training process is to minimize the KL-divergence of two probability distributions. After replacing d(· , ·) with KL-divergence and omitting some constants, the loss function is:

$$O_1 = -\sum_{(i,j)\in E} w_{ij} \log p_1(v_i, v_j) \tag{3}$$

The second-order proximity between a pair of vertices (u, v) in a network is the similarity between their neighborhood network structures. Mathematically, let $p_u = (w_{u,1}, \ldots, w_{u,|V|})$ denote the first-order proximity of u with all the other vertices, then the second-order proximity between u and v is determined by the similarity between p_u and p_v. If no vertex is linked from/to both u and v, the second-order proximity between u and v is 0. For each directed edge (i, j), the probability of "context" v_j generated by vertex v_i can be defined as:

$$p_2\left(v_j \mid v_i\right) = \frac{\exp\left(\overrightarrow{u}_j'^T \cdot \overrightarrow{u}_i\right)}{\sum_{k=1}^{|V|} \exp\left(\overrightarrow{u}_k'^T \cdot \overrightarrow{u}_i\right)} \tag{4}$$

where $|V|$ is the number of vertices or "contexts." $\overrightarrow{u_i}$ is the representation of v_i when it is treated as a vertex. $\overrightarrow{u_i}'$ is the representation of v_i when it is treated as a specific "context." The empirical distribution is $\hat{p}_2(\cdot\,|v_i)$. So, the objective function is as follows:

$$O_2 = \sum_{i\in V} \lambda_i d(\hat{p}_2\left(\cdot|v_i\right), p_2\left(\cdot|v_i\right)) \tag{5}$$

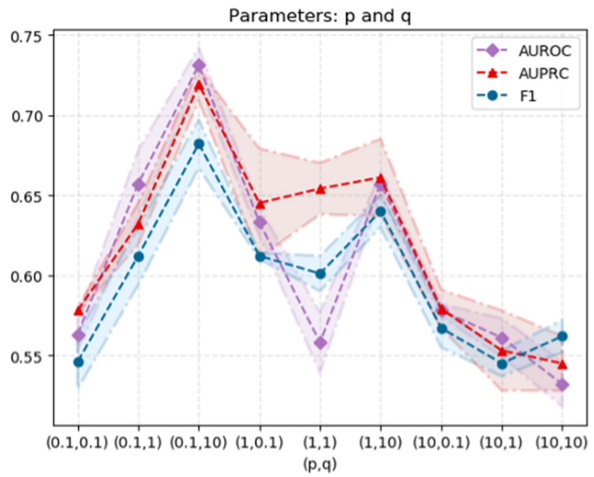

FIGURE 5 | The model performance varies as the hyperparameters (*p*, *q*) change in node2vec.

λ_i in the objective function represents the prestige of vertex*i* in the network, which can be measured by the degree or estimated through algorithms such as PageRank. The empirical distribution $\hat{p}_2(\cdot|v_i)$ is defined as $\hat{p}_2\left(v_j \mid v_i\right) = \frac{w_{ij}}{d_i}$, where w_{ij} is the weight of the edge (i, j) and d_i is the out-degree of vertex i, i.e., $d_i = \sum_{k\in N(i)} w_{ik}$, where $N(i)$ is the set of out-neighbors of v_i. After replacing d(· , ·) with KL-divergence, setting $\lambda_i = d_i$ and omitting some constants, the loss function is:

$$O_2 = -\sum_{(i,j)\in E} w_{ij} \log p_2\left(v_j|v_i\right) \tag{6}$$

The method in this paper is to train the LINE model which preserves the first-order proximity and second-order proximity separately and then concatenate the embeddings trained by the two methods for each vertex.

Reducing Feature Dimensions Using a Stacked Autoencoder

An autoencoder is an unsupervised model which is well known for its function of extracting features and reducing dimensionality. Aiming at minimizing the reconstruction errors between input and output, an autoencoder consists of two main parts, an encoder and a decoder. The hidden layer encoded features are the final low-dimensional output that plays a vital role in the downstream tasks. If the input node vector is x, the reconstructed node vector can be represented as $z(x) = g(w' \cdot f\left(w \cdot \mathrm{x}b\right) b')$, where f and g are active functions, w, $w'$are weights, and b, b' are biases. Hence, the objective function can be represented as Eq. 7, where represents the parameters, and L represents the loss function.

$$\theta = \underset{\theta}{\operatorname{argmin}} L\left(X, Z\right) \tag{7}$$

The SAE is a neural network composed of a multi-layer sparse autoencoder, which is used to boost performance of deep networks, and its structure is shown in **Figure 2C**. In SAE, the output of the previous layer of autoencoder is used as the input of the next layer of autoencoder. There are three steps to train a SAE. Firstly, a sparse autoencoder is trained on raw input and the trained sparse autoencoder is used to transform the raw input into a feature vector. Secondly, it uses the output of the former layer as input for the subsequent layer and repeats this process until the end of the training. Thirdly, after all the hidden layers are trained, back propagation algorithm is used to minimize the cost function and the pre-trained neural network can be fine-tuned with a labeled training set. SAE has achieved effective outcomes in many areas to extract feature vectors and reduce dimensionality. Alongside this trend, we enroll the SAE model in this proceeding for more impressive performance of predicting IS disease-related genes.

TABLE 1 | Top 10 genes predicted associated with ischemic stroke.

Gene ID	Gene name	Gene description	Score
51181	DCXR	Dicarbonyl and L-xylulose reductase	0.9854
22953	P2RX2	Purinergic receptor P2X 2	0.9762
57104	PNPLA2	Patatin like phospholipase domain containing 2	0.9723
3766	KCNJ10	Potassium inwardly rectifying channel subfamily J member 10	0.9645
3955	LFNG	LFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase	0.9631
10382	TUBB4A	Tubulin beta 4A class IVa	0.9543
2261	FGFR3	Fibroblast growth factor receptor 3	0.9532
84126	ATRIP	ATR interacting protein	0.9451
2182	ACSL4	Acyl-CoA synthetase long chain family member 4	0.9435
57511	COG6	Component of oligomeric Golgi complex 6	0.9410

Predicting Genes Associated With IS Using SVM

After low-dimensional gene features are generated, the SVM algorithm is trained to predict the disease-related genes of IS. The process of predicting such genes is considered as a node classification task. SVM has gained plenty of affirmations for its stability, simplicity, and effectiveness in the way of classification task. Therefore, SVM is engaged in our model analysis. We treat disease-related genes of IS as positive samples, then from the PPI network we randomly designate unlabeled genes of equivalent size as negative samples.

We use five-fold cross validation to evaluate the performance of the SVM classifier in the task of predicting IS disease-related genes. During the experiments, the standard Gaussian kernel is selected for performing the SVM classifier. Besides, we use the grid search method to select the optimal hyper-parameters.

RESULTS

Datasets

During the experiments, we downloaded two datasets, the disease-related genes of IS and the PPI network from public

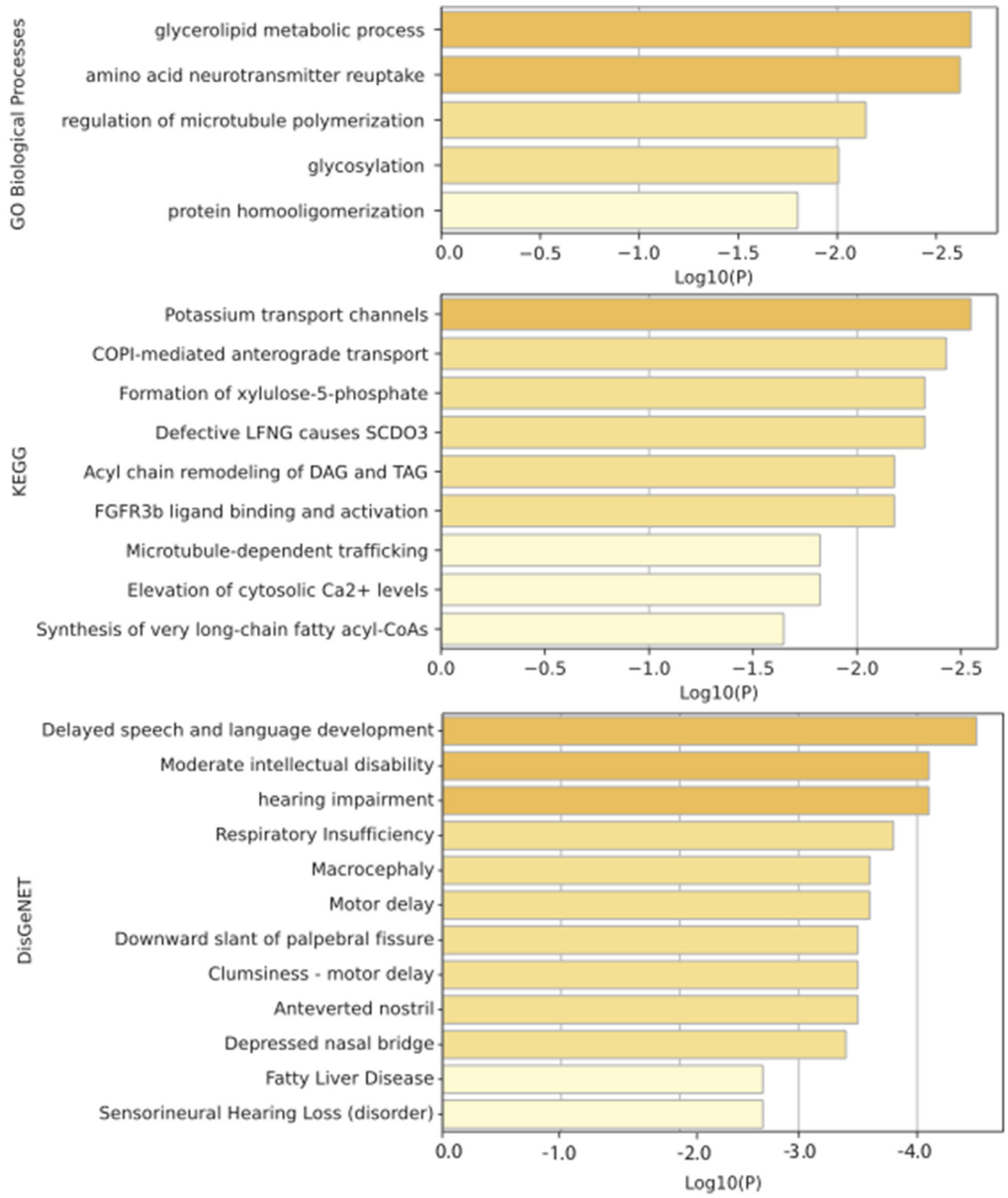

FIGURE 6 | Enrichment analysis of top 10 predicted IS-related genes based on GO, KEGG, and DisGeNet.

resources. The PPI network is originated from Menche et al. (2015), including 13,460 nodes and 141,296 edges. The genes associated with IS were downloaded from the DisGeNet database.[1] After analyzing and classifying corresponding genes related to IS or cerebral infarction as stated, we finally obtained 1195 IS-related genes.

Impact of Feature Dimensions on Predicting Performance

In order to explore the optimized dimension of NRL-based algorithms for predicting the disease-related genes of IS, we evaluated the performance of three NRL-based algorithms, i.e., DeepWalk, LINE, and Node2vec, using multiple levels of feature dimensions. Specifically, we run these NRL algorithms to generate features vectors in different dimension-levels, including 64, 128, 256, and 512. All features will be further processed by autoencoder to reduce noise; afterward, the autoencoder will output features in 64 dimensions for downstream predicting tasks. We compared their performance using five-fold cross validation; the results are presented in **Figure 3**.

We used area under the ROC curve (AUROC), area under the PR curve (AUPRC), and F1 scores to evaluate the performance of deepwalk, LINE, and node2vec in predicting IS-related genes using various feature dimensions. For LINE, the prediction

[1]https://www.disgenet.org/browser/0/1/1/C0026769

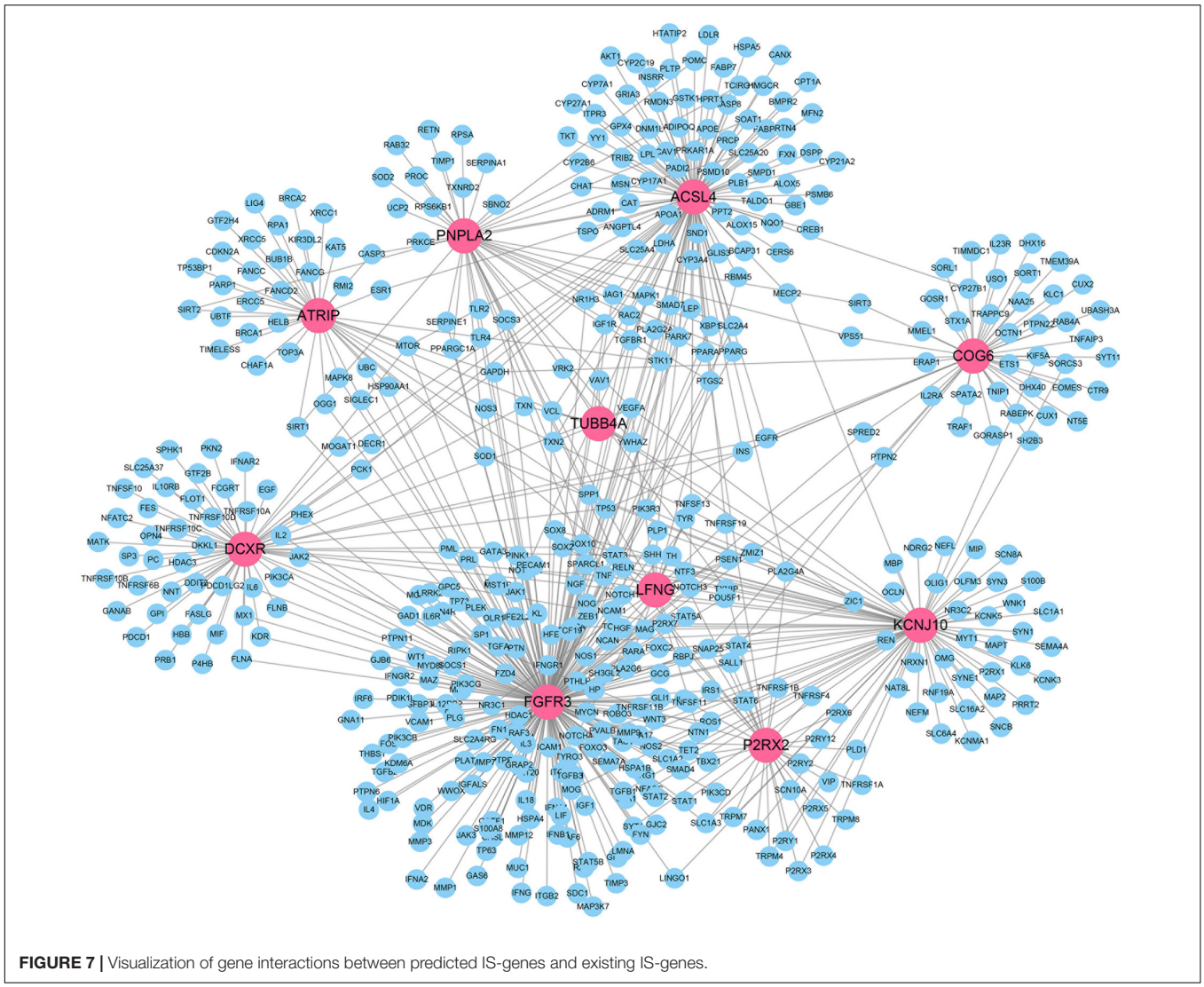

FIGURE 7 | Visualization of gene interactions between predicted IS-genes and existing IS-genes.

performance drops gradually as the feature dimension increases. For DeepWalk, the prediction performance drops from dim-64 to dim-256, while it increases when feature dimension is up to 512. For node2vec, the best performance is achieved at dim-64 and much better than the other two methods, while other feature dimensions achieve average performance.

For intuitional comparison, we summarized the best performance of these three algorithms as shown in **Figure 4**. We can see that Node2vec with dim-64 provides the most effective outcomes. Therefore, in the final predicting model, we adopt node2vec to learn the graph embedding with 64 feature dimensions.

Effects of Hyper-Parameters on Ischemic Stroke-Related Gene Prediction

As mentioned above, the computational workflow use node2vec to capture the topological structure information from the PPI network, followed by extracting low-dimensional features, and predicting disease-related genes based on the SVM classifier. It has been shown in relative researches that the hyper-parameters used in node2vec have considerable impact on the prediction performance. In order to explore the optimized hyper-parameters, we performed a grid search for the hyper-parameters of node2vec, namely p and q, to test the performance. We randomly select parameters $p \in \{0.1, 1, 10\}$ and $q \in \{0.1, 1, 10\}$. When p is relatively small, the random walk is more inclined to visit the nodes that have been visited. When $q > 1$, the random walk is biased to BFS, and when $q < 1$, the random walk clings to DFS. The standard deviation of 50% cross validation and the results are shown in **Figure 5**.

From the data, when $p = 0.1$ and $q = 10$, the AUROC value of the node2vec algorithm achieves its maximum (0.731), which elucidates the optimized choice of hyper-parameters.

Top Genes Related to Ischemic Stroke

In order to verify the performance of the algorithm in predicting novel genes related to IS, we use existing all-known genes related to IS as the training set and the unknown genes as the test set.

Then we rank the probability of final prediction. We select the top 10 genes and list their gene ID and name in **Table 1**.

Recent studies have shown the correlation between these discovered genes and IS. Cui et al. (2021) utilized lentivirus *in vitro* infection and *in vivo* administration methods to prove that knockdown of ACSL4 alleviated brain injury after IS. Zhao et al. (2020) performed real-time polymerase chain reaction (PCR) to analyze the association between PNPLA2 rs1138693 (T > C) genotype and the risk of IS. Wang J.F. et al. (2020) proved P2RX2 as an up-regulated gene in myocardial infarction using gene ontology (GO) analysis and pathway enrichment analysis in a comparative study of gene expression profiles rooted in acute ischemia and infarction.

Functional Analysis of the Top Predicted IS-Genes

We performed enrichment analysis for the top 10 IS-genes predicted by our method based on GO, KEGG, and DisGeNet, and the results are illustrated in the **Figure 6**. The most GO biological process enriched is the glycerolipid metabolic process. Wang et al. (2021) has proved that the glycerophospholipid metabolism plays a role in IS. KEGG analysis revealed the importance of potassium transport channels in IS, and this also was demonstrated in the work of Chen et al. (2016), where they found that potassium channels can be a potential pharmacological target for IS to slow down cerebral edema formation. The enrichment results from DisGeNet show that the top 10 IS-related genes we predicted are related to language development, intellectual disability, hearing impairment, and motor delays, and these symptoms happen a lot in clinic after occurring IS.

We also visualized the gene network between the top 10 predicted IS-genes and the known IS = related genes from DisGeNet in **Figure 7**. We can see that the top 10 genes predicted by our method are closely connected to the known IS-genes. The gene with highest degree is FGFR3, and the fibroblast growth factors have shown great therapeutic potential in treatment of IS.

CONCLUSION

It is quite crucial to discover the disease-related genes of IS for future medical treatment and more accurate diagnosis. In this paper, we utilize NRL methods for the task of identifying disease-related genes and test the novel NRL-based framework to discover IS-related genes. There are three main components in the whole operating process: capturing the global topological information of the PPI, utilizing a SAE to represent vectors into low-dimensional feature space, and training an SVM classifier to predict disease-related genes. The experimental results show that the proposed NRL-based algorithm could achieve considerable accuracy in predicting the genes of IS. Furthermore, the introduced NRL-based algorithms are exploiting and stable to be forwarded to many other fields of potential gene prediction.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HLiu, LH, and SW conceived the study. HLiu designed and performed the experiments. HLiu, SX, HLi, JG, ZW, and XL analyzed the data and wrote and revised the manuscript. SW supervised the study. All authors contributed to the article and approved the submitted version.

REFERENCES

Cao, M., Pietras, C. M., Feng, X., Doroschak, K. J., Schaffner, T., Park, J., et al. (2014). New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* 30, i219–i227.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Chen, Y.-J., Nguyen, H. M., O'Donnell, M. E., and Wulff, H. (2016). Potassium channels in ischemic stroke. *FASEB J.* 30, 1224.19–1224.19.

Cheng, Y.-C., Cole, J. W., Kittner, S. J., and Mitchell, B. D. (2014). Genetics of ischemic stroke in young adults. *Circ. Cardiovasc. Genet.* 7, 383–392. doi: 10.1161/circgenetics.113.000390

Cui, Y., Zhang, Y., Zhao, X., Shao, L., Liu, G., Sun, C., et al. (2021). ACSL4 exacerbates ischemic stroke by promoting ferroptosis-induced brain injury and neuroinflammation. *Brain. Behav. Immun.* 93, 312–321. doi: 10.1016/j.bbi.2021.01.003

Dai, J., Ren, J., and Du, W. (2020). Decomposition-based Bayesian network structure learning algorithm using local topology information. *Knowl. Based Syst.* 195:105602. doi: 10.1016/j.knosys.2020.105602

Embar, V., Handen, A., and Ganapathiraju, M. K. (2016). Is the average shortest path length of gene set a reflection of their biological relatedness? *J. Bioinform. Comput. Biol.* 14, 41–42.

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 855–864.

Jian, T., Meng, Q., Wang, M., Ming, Z., and Mei, Q. (2015). "LINE: Large-scale Information Network Embedding," in *Proceedings of the International World Wide Web Conferences Steering Committee*, (Geneva: International World Wide Web Conferences Steering Committee).

Kumar, A. A., Van Laer, L., Alaerts, M., Ardeshirdavani, A., Moreau, Y., Laukens, K., et al. (2018). pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. *Bioinformatics* 34, 2254–2262. doi: 10.1093/bioinformatics/bty079

Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. (2014). Exploring strategies for training deep neural networks. *J. Mach. Learn. Res.* 1, 1–40.

Matarin, M., Singleton, A., Hardy, J., and Meschia, J. (2010). The genetics of ischaemic stroke. *J. Intern. Med.* 267, 139–155. doi: 10.1111/j.1365-2796.2009.02202.x

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Molet, M., Stagner, J. P., and Miller, H. C. (2013). Guilt by association and honor by association: The role of acquired equivalence. *Psychon. Bull. Rev.* 20, 385–390. doi: 10.3758/s13423-012-0346-3

Nguyen, T. P., and Ho, T. B. (2012). Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. *Artif. Intell. Med.* 54, 63–71. doi: 10.1016/j.artmed.2011.09.003

Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein–protein interactions. *J. Med. Genet.* 43, 691–698.

Peng, J., Guan, J., Hui, W., and Shang, X. (2021a). A novel subnetwork representation learning method for uncovering disease-disease relationships. *Methods* 192, 77–84. doi: 10.1016/j.ymeth.2020.09.002

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019). A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* 35, 4364–4371. doi: 10.1093/bioinformatics/btz254

Peng, J., Wang, T., Hu, J., Wang, Y., and Chen, J. (2016). Constructing networks of organelle functional modules in *Arabidopsis*. *Curr. Genomics* 17, 427–438. doi: 10.2174/1389202917666160726151048

Peng, J., Wang, Y., Guan, J., Li, J., Han, R., Hao, J., et al. (2021b). An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Brief. Bioinform.* doi: 10.1093/bib/bbaa430 [Epub ahead of print].

Peng, J., Xue, H., Wei, Z., Tuncali, I., Hao, J., and Shang, X. (2021c). Integrating multi-network topology for gene function prediction using deep neural networks. *Brief. Bioinform.* 22, 2096–2105. doi: 10.1093/bib/bbaa036

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, 701–710.

Sacco, R. L., Kasner, S. E., Broderick, J. P., Caplan, L. R., Connors, J. J., Culebras, A., et al. (2013). An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 44, 2064–2089.

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., and Baudot, A. (2017). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 35, 497–505. doi: 10.1093/bioinformatics/bty637

Vuillon, L., and Lesieur, C. (2015). From local to global changes in proteins: a network view. *Curr. Opin. Struct. Biol.* 31, 1–8. doi: 10.1016/j.sbi.2015.02.015

Wang, J.-F., Huang, Y., Lu, S. F., Hong, H., Xu, S. J., Xie, J. S., et al. (2020). Comparative study of gene expression profiles rooted in acute myocardial infarction and ischemic/reperfusion rat models. *Am. J. Cardiovasc. Dis.* 10:84.

Wang, T., Peng, J., Peng, Q., Wang, Y., and Chen, J. F. S. M. (2019a). Fast and scalable network motif discovery for exploring higher-order network organizations. *Methods* 173, 83–93. doi: 10.1016/j.ymeth.2019.07.008

Wang, T., Peng, Q., Liu, B., Liu, X., Liu, Y., Peng, J., et al. (2019b). eQTLMAPT: fast and accurate eQTL mediation analysis with efficient permutation testing approaches. *Front. Genet.* 10:1309. doi: 10.3389/fgene.2019.01309

Wang, T., Peng, Q., Liu, B., Liu, Y., and Wang, Y. (2020). Disease module identification based on representation learning of complex networks integrated from GWAS, eQTL summaries, and human interactome. *Front. Bioeng. Biotechnol.* 8:418. doi: 10.3389/fbioe.2020.00418

Wang, X., Zhang, L., Sun, W., Pei, L. L., Tian, M., Liang, J., et al. (2021). Changes of metabolites in acute ischemic stroke and its subtypes. *Front. Neurosci.* 14:580929. doi: 10.3389/fnins.2020.580929

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22, 2800–2805. doi: 10.1093/bioinformatics/btl467

Yang, W., Han, J., Ma, J., Feng, Y., Hou, Q., Wang, Z., et al. (2019). Prediction of key gene function in spinal muscular atrophy using guilt by association method based on network and gene ontology. *Exp. Ther. Med.* 17, 2561–2566.

Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.* 6:252ra123. doi: 10.1126/scitranslmed.3009262

Zhao, H., Haojun, Z., Tingting, Z., Meihua, Y., Xi, D., Liang, M., et al. (2020). Association between the polymorphism of PNPLA2 gene and the risk of ischemic stroke in type 2 diabetic patients in Chinese Han Population. *Chin. J. Clin. Pharmacol. Ther.* 25:664.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with one of the authors XL.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 September 2021
doi: 10.3389/fgene.2021.739677

AdImpute: An Imputation Method for Single-Cell RNA-Seq Data Based on Semi-Supervised Autoencoders

*Li Xu[1,2], Yin Xu[3], Tong Xue[1], Xinyu Zhang[1] and Jin Li[1]**

[1] College of Computer Science and Technology, Harbin Engineering University, Harbin, China, [2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China, [3] School of Mathematics, Harbin Institute of Technology, Harbin, China

Motivation: The emergence of single-cell RNA sequencing (scRNA-seq) technology has paved the way for measuring RNA levels at single-cell resolution to study precise biological functions. However, the presence of a large number of missing values in its data will affect downstream analysis. This paper presents AdImpute: an imputation method based on semi-supervised autoencoders. The method uses another imputation method (DrImpute is used as an example) to fill the results as imputation weights of the autoencoder, and applies the cost function with imputation weights to learn the latent information in the data to achieve more accurate imputation.

Results: As shown in clustering experiments with the simulated data sets and the real data sets, AdImpute is more accurate than other four publicly available scRNA-seq imputation methods, and minimally modifies the biologically silent genes. Overall, AdImpute is an accurate and robust imputation method.

Keywords: scRNA-seq, missing value filling, semi-supervised learning, autoencoder, imputation method

OPEN ACCESS

Edited by:
Jiajie Peng, Northwestern Polytechnical University, China

Reviewed by:
Weihua Guan, University of Minnesota Twin Cities, United States
Fukang Zhu, Jilin University, China

****Correspondence:***
Jin Li
lijinokok@hrbeu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *11 July 2021*
Accepted: *18 August 2021*
Published: *08 September 2021*

Citation:
Xu L, Xu Y, Xue T, Zhang X and Li J (2021) AdImpute: An Imputation Method for Single-Cell RNA-Seq Data Based on Semi-Supervised Autoencoders. Front. Genet. 12:739677. doi: 10.3389/fgene.2021.739677

INTRODUCTION

With the development of high-throughput sequencing technology, the emergence of single-cell RNA sequencing (scRNA-seq) technology in genomic sequencing has become a hot topic in recent years (Wagner et al., 2016; Kalisky et al., 2018). Compared with bulk RNA sequencing sequences, scRNA sequences have a relatively high noise level, especially due to so-called dropouts (Vallejos et al., 2015; Lun et al., 2016; Ziegenhain et al., 2017). Dropouts are a special type of missing values due to low RNA input in sequencing experiments and the randomness of gene expression patterns at the single cell level. The presence of dropouts often misleads downstream analysis, such as data visualization, cell clustering, and differential expression analysis (Stegle et al., 2015; Bacher and Kendziorski, 2016; Svensson et al., 2017).

Based on different principles, a variety of single cell RNA-seq data imputation methods have been proposed (Chen and Zhou, 2018; Huang et al., 2018; Van Dijk et al., 2018; Eraslan et al., 2019; Hu et al., 2020; Qi et al., 2021). ScImpute (Li and Li, 2018) divides genes into two groups based on dropout probability (unreliable and reliable classification: A_j, B_j), and the dropout probability is

estimated by a mixed model. Scientific computing estimates A_j by processing B_j as gold standard data. In the first version, a weighted lasso model is used to find similar cells among other cells in B_j genes. Then use the linear regression model of the most similar unit as the estimate of A_j. DrImpute (Gong et al., 2018) is an integrated method, which is designed based on the consistent clustering results of scRNA-seq data. In other words, it performs multiple clusters and imputes based on the average of similar cell expression. AutoImpute (Talwar et al., 2018) is a method of imputing dropouts based on an autoencoder. It uses over-complete autoencoders to capture the distribution of given sparse gene expression data, and regenerates complete expression data. DeepImpute (Zhang and Zhang, 2020) is an imputation method based on deep neural networks. The method uses missing layers and loss functions to learn patterns in the data to achieve accurate imputation.

At present, machine learning methods are increasingly used in bioinformatics, and many achievements have been made (Peng et al., 2021a,b). We have conducted a lot of clustering experiments on the existing imputation method. According to the experimental results, we found that the machine learning methods did not perform well. The analysis revealed two reasons. One is a large number of zeros in the raw data, making it difficult for machine learning methods to extract deep information from the data. Instead, most of the zeros are regarded as true zeros, that is, no padding is performed, so the data filled by the machine learning method is more discrete. The second reason is that after using some deep learning-based missing value filling methods to fill in, the output data contains negative values, but the actual gene expression values should all be non-negative values. Based on this, we propose an imputation method AdImpute (**Figure 1**) based on a semi-supervised autoencoder, which combines ordinary imputation methods with machine learning methods to better implement imputation.

An Autoencoder is a type of artificial neural network used in semi-supervised learning and unsupervised learning. Its function is to perform representation learning on the input information by using the input information as the learning target. A number of recent studies describe applications of autoencoders in molecular biology.

In order to solve the problem of difficult to extract the deep information of the data, AdImpute introduces a set

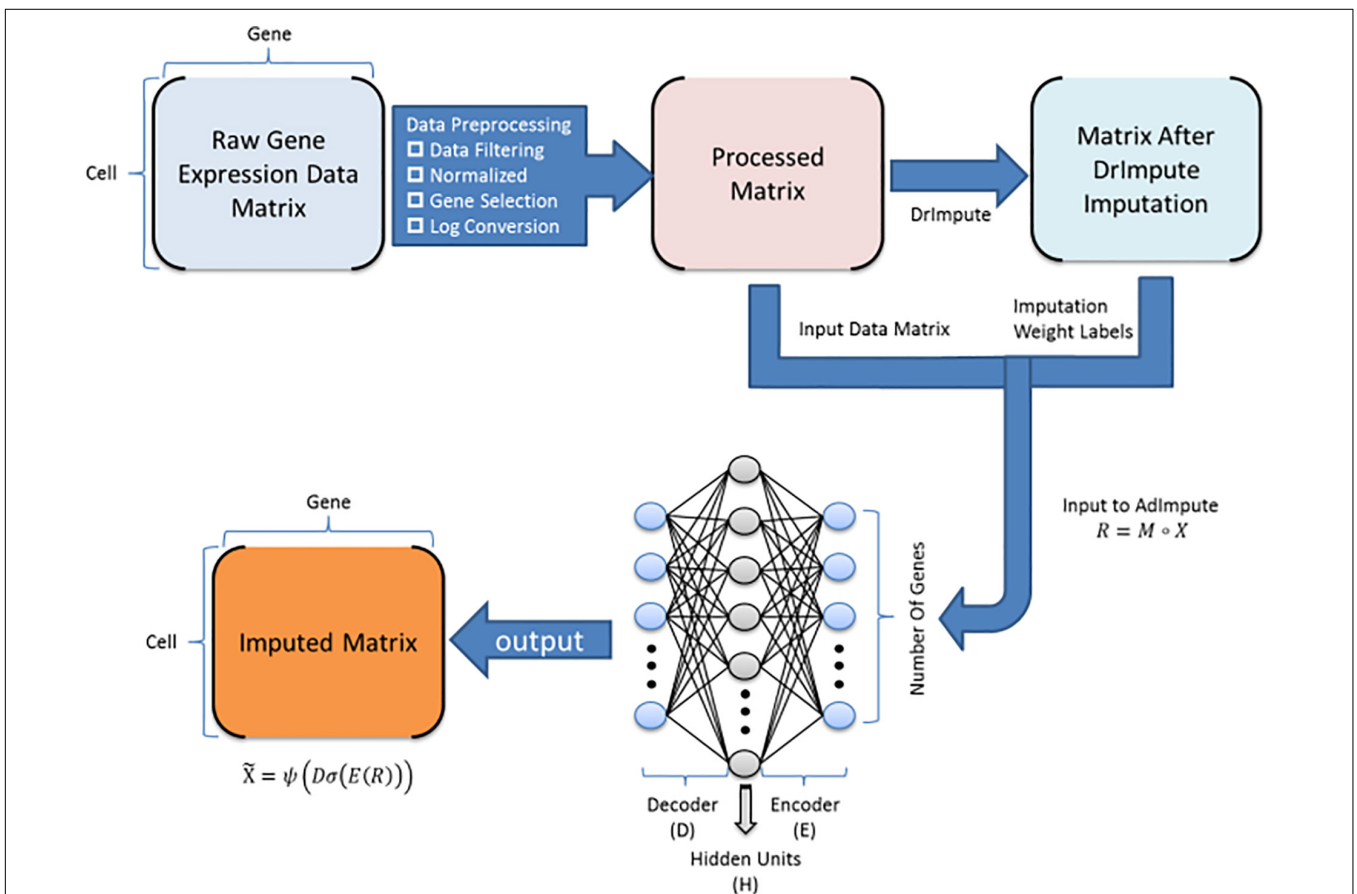

FIGURE 1 | AdImpute pipeline: the pre-processing stage of AdImpute requires screening of raw gene expression data, normalizing by library size, and pruning through gene selection and logarithmic transformation. Afterward, AdImpute first fills the pre-processed matrix with DrImpute, and uses the result of DrImpute as an imputation weight label. Then the label is input into the AdImpute model together with the pre-processed matrix to learn gene expression data. Finally, the missing data value filling and the input matrix reconstruction are done.

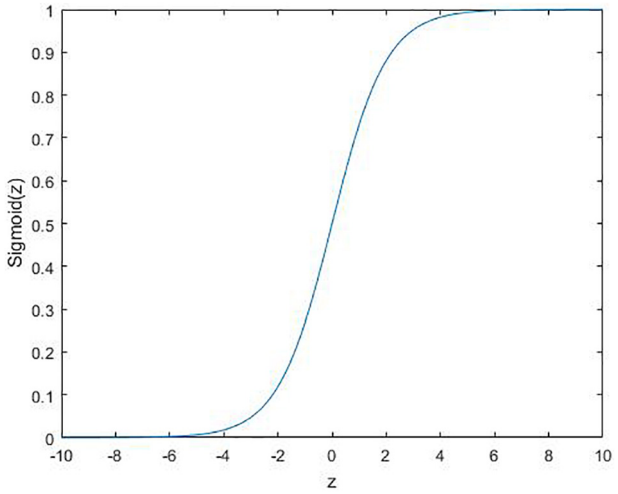

FIGURE 2 | Sigmoid function: $f(z) = \frac{1}{1+e^{-z}}$.

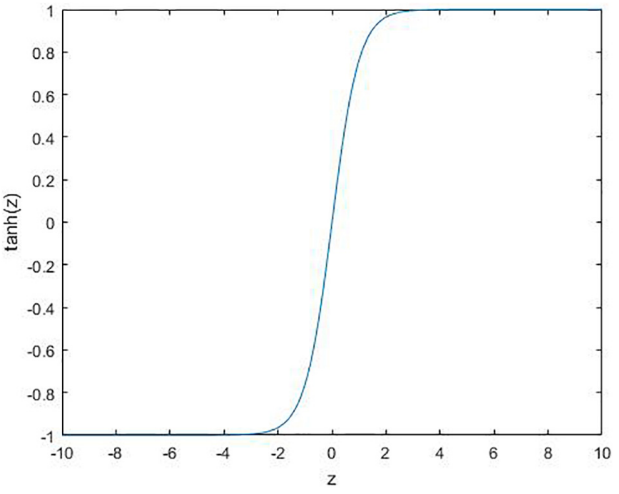

FIGURE 3 | Tanh function: $\tanh(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$.

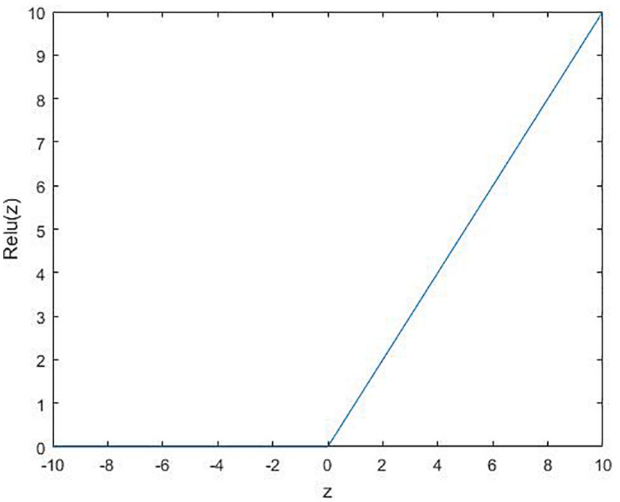

FIGURE 4 | Relu function: $\text{relu}(z) = \max(0, z)$.

of data imputed by DrImpute as an imputed weight label (DrImpute method can be replaced, this article selects the current mainstream method, if there is a better one, you can replace it). While using the autoencoder to impute dropouts, AdImpute adds an imputation weight term to the cost function and compares it with the label data. For a zero value that may be a missing value, the larger the label data value is, the more likely it is to be a missing value, so as to achieve semi-supervised learning. We also give a Relu activation function to the decoding layer to solve the situation of negative values in the filled data.

An example is given to better understand the principle of this method. If we compare the imputation process to an exam, then the unsupervised machine learning method is to complete a test paper normally, and supervised machine learning is to complete a test paper under the premise of having a standard answer. Semi-supervision is equivalent to finding a test paper of a student with good grades as a reference to complete my test paper.

In reality, supervision is meaningless for imputation. Current machine learning algorithms are all based on unsupervised. Here, we first proposed the idea of applying semi-supervision to imputation, and verified the superiority with the help of clustering results.

MATERIALS AND METHODS

Autoencoder

In simple terms, the autoencoder is the process of reducing the dimension after encoding the raw data, so as to discover the hidden rules among the data. The autoencoder is composed of encoder E and decoder D. The encoder first maps the input data X to the latent space H:

$$H = \phi(EX) \tag{1}$$

where ϕ is the activation function. Several commonly used functions are shown in **Figures 2–4**.

In the training phase, the encoder and decoder are usually learned by minimizing the Euclidean cost function:

$$\arg\min_{D,E} ||\mathrm{X} - \mathrm{D}\phi(EX)||_F^2 \tag{2}$$

There are several variants of the autoencoder model: multi-layer autoencoder and regularized autoencoder. The multi-layer autoencoder is created by nesting the autoencoder inside another autoencoder. Mathematically, this is expressed as:

$$\arg\min_{D's,E's} ||\mathrm{X} - \mathrm{D}_1\phi(D_2 \cdots \phi D_N(\phi(E_N \cdots \phi E_1(X) \cdots)||_F^2 \tag{3}$$

The cost function used by the regularized autoencoder can encourage the model to learn other features, rather than copying

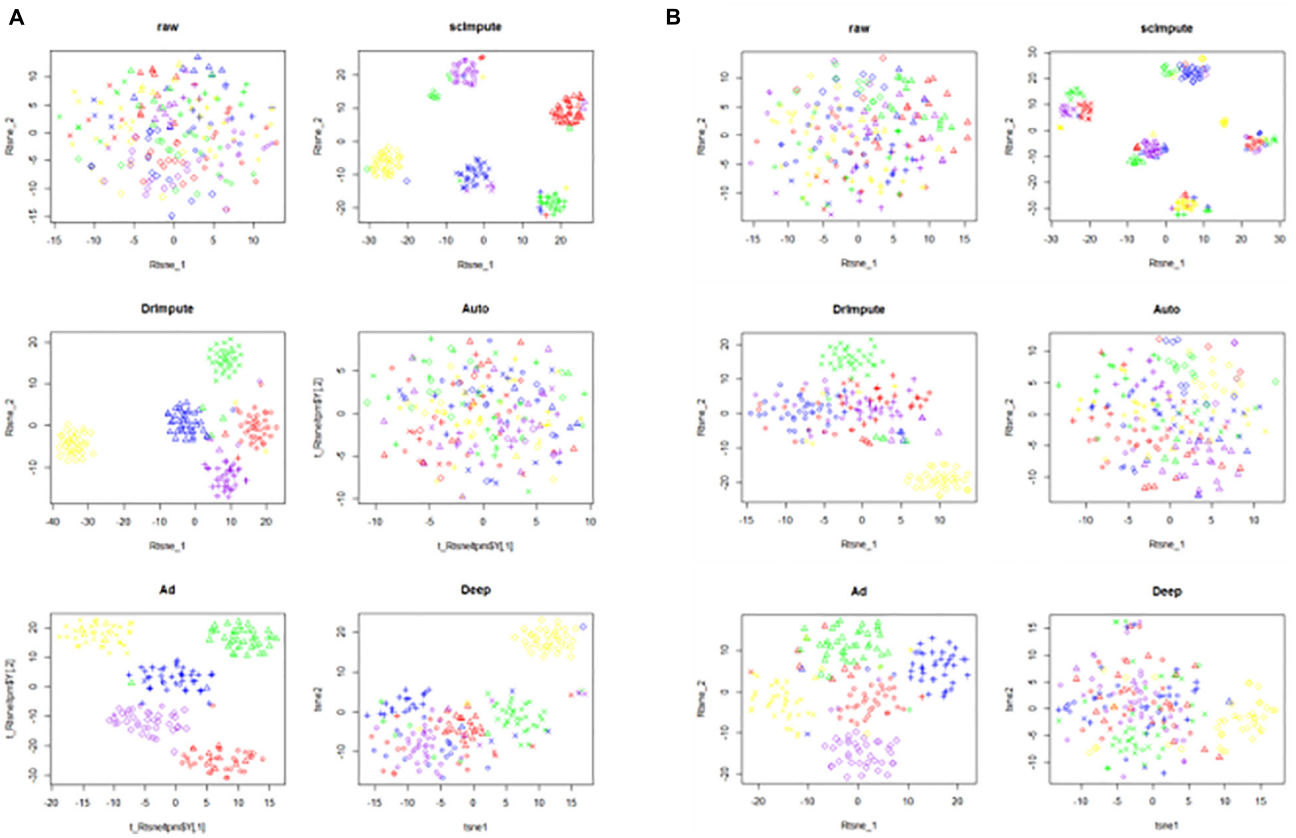

FIGURE 5 | The visualization results of clustering on the simulated data sets. The six images are the clustering results of the raw data and the clustering results imputed by scImpute, DrImpute, AutoImpute, AdImpute, and DeepImpute. And **(A)** is the clustering visualization results on simu1 data set; **(B)** is the clustering visualization results on simu2 data set.

the input to the output. These characteristics include sparse representation, robustness to noise or missing inputs, etc. Even if the model capacity is large enough to learn a meaningless identity function, the nonlinear and over-complete regular autoencoder can still learn some useful information about the data distribution from the data. The regularized autoencoder can be expressed as follows:

$$\arg\min_{D,E} ||\mathrm{X} - \mathrm{D}\phi(EX)||_F^2 + \lambda\Re(E, D) \tag{4}$$

where λ is the regularization coefficient, and the regularizer R is a real function about E and D.

The Design and Implementation of AdImpute

AdImpute is a missing value filling method based on semi-supervised autoencoder. While using a complete autoencoder to capture the distribution of the given sparse gene expression data, AdImpute introduces the data filled by DrImpute as the imputation weight label of the model, which makes the regenerated complete expression data obtain higher quality.

The purpose of AdImpute is to estimate these dropouts by looking for the full version of gene expression data. The model of the measured value is:

$$R = M \circ X \tag{5}$$

where ° is the Hadamard product, M is a binary mask containing 1, R contains a non-zero term, and elsewhere is 0. X represents the count matrix to be estimated.

AdImpute needs to import the data filled by DrImpute into the model as the imputation weight label of the model, which is recorded as F. Then the sparse gene expression matrix $M^{\circ}X$ is input into the autoencoder, and it is trained to learn the best encoder and decoder functions by minimizing the cost function. In order to prevent the overfitting of non-zero values in the count matrix, we regularize the learned encoder and decoder matrices. The cost function is as follows:

$$\min_{D,E} ||\mathrm{R} - \mathrm{D}\sigma(E(R))||_O^2 + \delta||\mathrm{F}\text{-}\mathrm{D}\sigma(E(R))||_O^2 + \frac{\lambda}{2}\delta\left(||E||_F^2 + ||D||_F^2\right) \tag{6}$$

where E is the encoder matrix, D is the decoder matrix, and λ is the regularization coefficient. In the formula (6), $\delta||\mathrm{F} - \mathrm{D}\sigma(E(R))||_O^2$ is the imputation weight item, F is the imputation weight label, and δ is the weight of the imputation

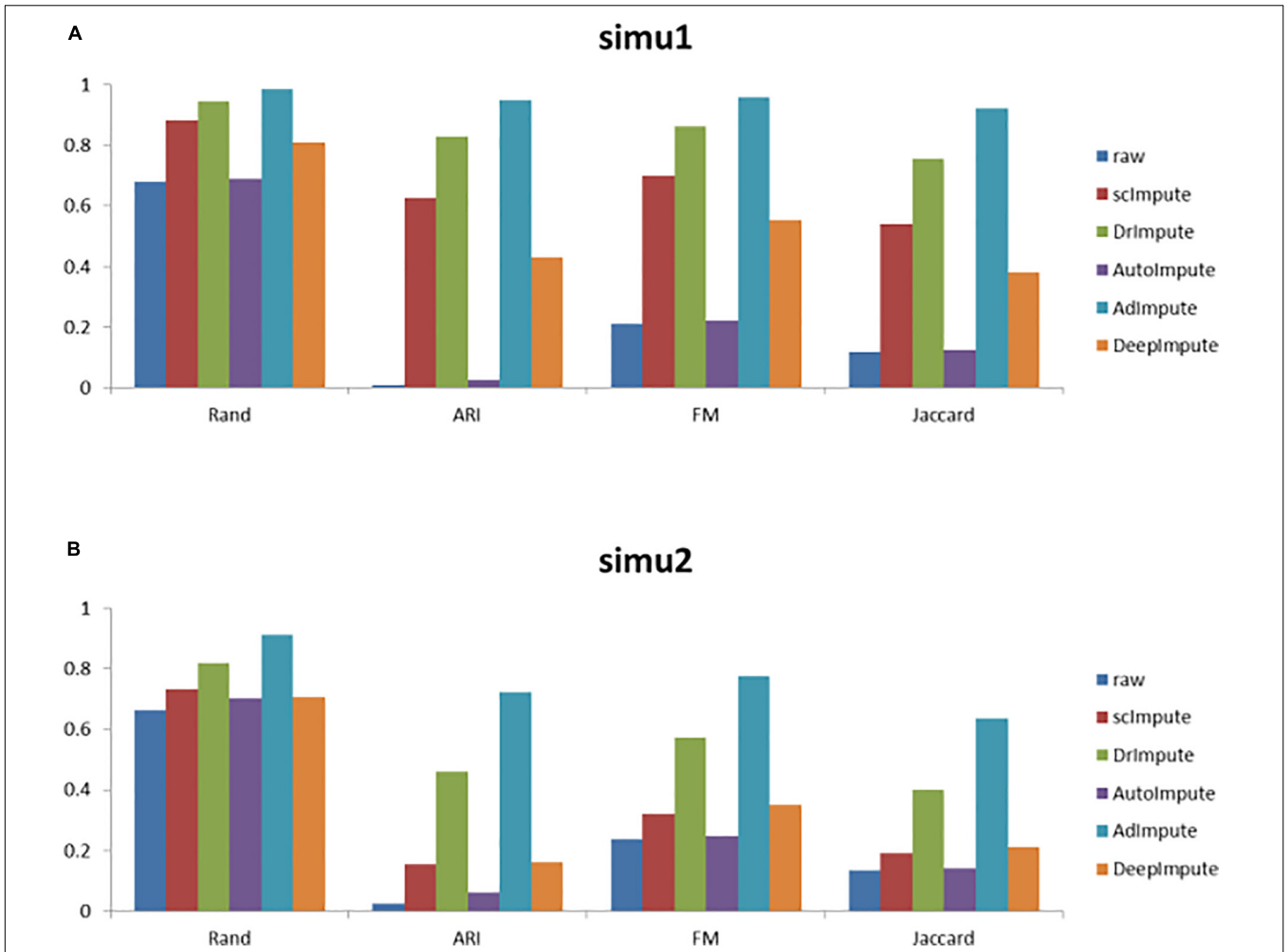

FIGURE 6 | The histogram of clustering evaluation indexes on the simulated data sets, including the clustering evaluation indexes of raw data and the clustering evaluation indexes after imputing by scImpute, DrImpute, AutoImpute, AdImpute, and DeepImpute. And **(A)** is the clustering evaluation indexes on simu1 data set; **(B)** is the clustering evaluation indexes on simu2 data set.

weight item. $||\cdot||_O$ means that the loss is calculated only for the non-zero counts present in the sparse expression matrix $M \circ X$, and σ is the Sigmoid activation function applied to the encoder layer in the neural network.

Finally, after the training and learning encoder and decoder matrix, the filled expression matrix is as follows:

$$\tilde{X} = \psi(D\sigma(E(R))) \quad (7)$$

where ψ is the Relu activation function applied to the decoder layer in the neural network.

The AdImpute model consists of a fully connected multi-layer perceptron with three layers: input layer, hidden layer and output layer. The model uses an imputation weight label composed of DrImpute-filled data to improve the missing value filling effect. The gradient is calculated by back propagation of the error, and the gradient descent method is used for training to reach the minimum value of the cost function (6). The RMSProp Optimizer is used to adjust the learning rate so as to avoid falling into a local minimum and reach the minimum of the cost function faster. Both the encoder matrix E and the decoder matrix D are subject to the initialization of random normal distribution. The output of the decoder uses Relu as the activation function.

The selection of hyper-parameters is as follows:

(1) Regularization coefficient λ is used to control the contribution of the regular term to the cost function.
(2) The weight δ of the imputation weight term is used to control the contribution of the imputation weight term to the cost function.

TABLE 1 | The ranking of clustering effects on the simulated data sets.

	scImpute	DrImpute	AutoImpute	AdImpute	DeepImpute
simu1	3	2	5	1	4
simu2	4	2	5	1	3

1 represents the best and 5 represents the worst in the table.

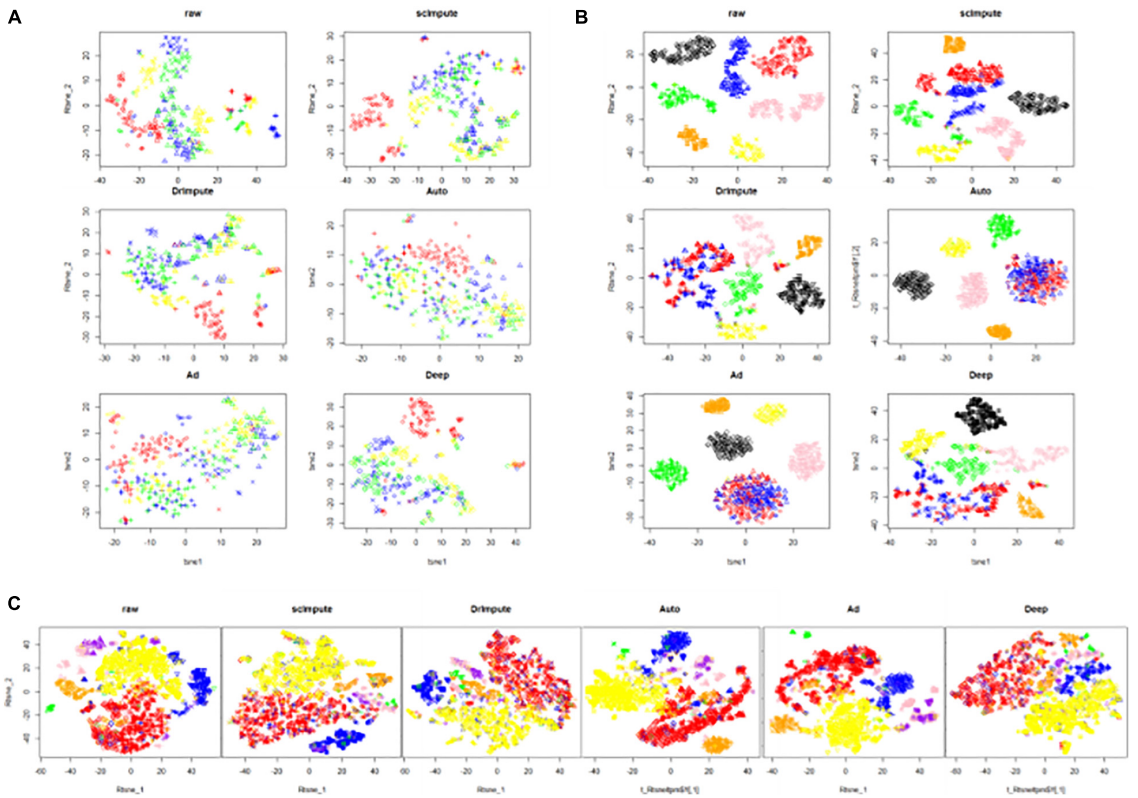

FIGURE 7 | The visualization results of clustering on the real data sets. The six images are the clustering results of the raw data and the clustering results imputed by scImpute, DrImpute, AutoImpute, AdImpute, and DeepImpute. And **(A)** is the clustering visualization results on Trapnell (GSE52529) data set; **(B)** is the clustering visualization results on hPSC (GSE75748) data set; **(C)** is the clustering visualization results on Romanov (GSE74672) data set.

(3) The size of the hidden layer (the dimension of the potential space).
(4) The initial value of the learning rate.
(5) Threshold. The change of the cost function value in successive iterations is less than the threshold value, which means convergence and stops the gradient descent.

RESULTS

A good imputation method can retain most of the real available information for the raw data. Therefore, in order to measure the quality of the missing value filling methods, we choose cluster analysis in the downstream analysis. We will select some data sets and use five methods to impute the dropouts, and use the results to perform cluster analysis.

By analyzing the results of the clustering, we estimated the advantages and disadvantages of the dropouts imputation methods. The cluster evaluation indicators used in this paper are rand, ARI, FM, and Jaccard.

The Clustering Experiment on the Simulated Data Sets

We use CIDR (Lin et al., 2017) to generate two simulated data sets simu1 and simu2. The details of simu1 and simu2 is provided in the section "Data availability." We label the generated data and mark the actual cell clustering results. After preprocessing the raw data, we use scImpute, DrImpute, AutoImpute, AdImpute, and DeepImpute to impute the dropouts. Based on the imputed data results, T-SNE for dimensionality reduction and visualization is carried out, and then K-means clustering is used. The visualization results of clustering are shown in **Figure 5**.

Based on the clustering results, we calculate the cluster evaluation indexes. The results are given by **Supplementary Table 1**. In order to analyze the experimental results more intuitively, we give a histogram of clustering evaluation indexes in **Figure 6**.

Analyzing the results of the above experiments, we can find that AdImpute has a very good performance in the clustering experiment on the simulated data sets. The performance of AutoImpute is not ideal, scImpute and DeepImpute are always slightly inferior than DrImpute. In general, AdImpute performs best on the simulated data sets. And the ranking of clustering effects is shown in **Table 1**.

The Clustering Experiment on the Real Data Sets

In the part, we select three real data sets: Trapnell (Trapnell et al., 2014), hPSC (Chu et al., 2016), and Romanov

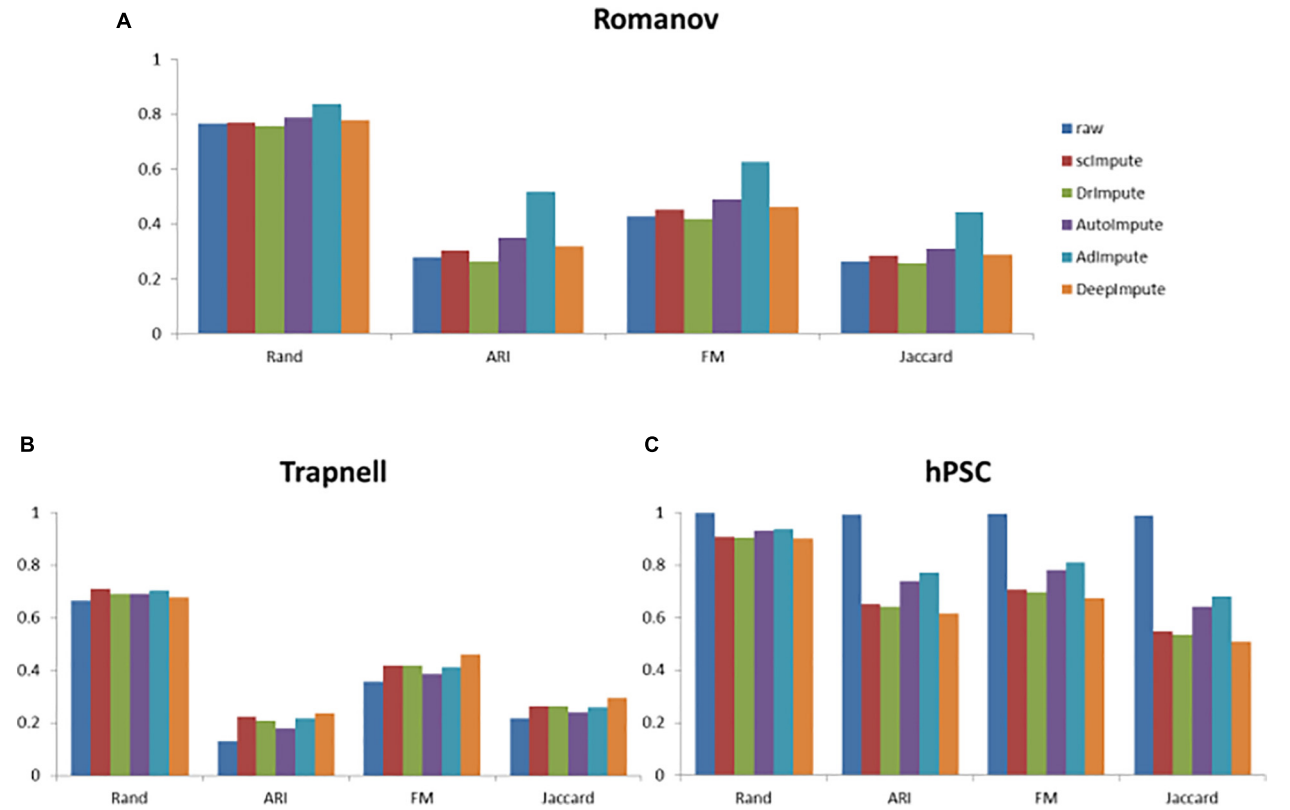

FIGURE 8 | The histogram of clustering evaluation indexes on the real data sets, including the clustering evaluation indexes of raw data and the clustering evaluation indexes after imputing by scImpute, DrImpute, AutoImpute, AdImpute, and DeepImpute. And **(A)** is the clustering evaluation indexes on Trapnell (GSE52529) data set; **(B)** is the clustering evaluation indexes on hPSC (GSE75748) data set; **(C)** is the clustering evaluation indexes on Romanov (GSE74672) data set.

(Romanov et al., 2017). After preprocessing the raw data, we use scImpute, DrImpute, AutoImpute, AdImpute, and DeepImpute to impute the dropouts. Based on the imputed data results, T-SNE for dimensionality reduction and visualization is carried out, and then K-means clustering is used. The visualization results of clustering are shown in **Figure 7**.

Based on the clustering results, we also calculate the cluster evaluation indexes. The results are given by **Supplementary Table 2**. In order to analyze the experimental results more intuitively, we give a histogram of clustering evaluation indexes in **Figure 8**.

Analyzing the experimental results, we can find that AdImpute still has a good performance in the clustering experiment on the real data sets. Despite being slightly inferior to scImpute in Trapnell data set, the overall performance is still the best among these methods. AutoImpute and DeepImpute do not perform well on the simulated data sets, but behave well on the real data sets. The performance of scImpute is unstable, and DrImpute is not ideal. Through the results on hPSC data set, we can see that AdImpute has minimally modified the expression of real biological silencing genes. Overall, AdImpute still performs best on the real data sets. And the ranking of clustering effects is shown in **Table 2**.

TABLE 2 | The ranking of clustering effects on the real data sets.

	scImpute	DrImpute	AutoImpute	AdImpute	DeepImpute
Trapnell	1	4	5	2	3
hPSC	3	4	2	1	5
Romanov	4	5	2	1	3

1 represents the best and 5 represents the worst in the table.

DISCUSSION

Since the scRNA-seq data suffers from dropout events that hinder the downstream analysis of data, we propose a statistical imputation method AdImpute to denoise the scRNA-seq data. AdImpute aims to implement data recovery and maintain the heterogeneity of gene expression across cells. One of the advantages of AdImpute is that it can be incorporated into most of the downstream analysis tools for the scRNA-seq data. In this paper, we perform downstream analysis experiments in the simulated datasets and real datasets, and the results show that our method improves the raw data and outperforms the other imputation methods.

Rand, ARI, FM, and Jaccard Index were used to measure the clustering results of imputed data. AdImpute performs well in the

clustering experiments of the simulated data sets and the real data sets. In the simulated data sets, it can be seen from **Figure 6** that the clustering results of AdImpute is significantly better than that of the other three algorithms when $v = 9/10$.

Because the data loss degree of real data is unknown, there may be a large number of true zeros, which can reflect the judgment ability of each algorithm to distinguish between missing zeros and true zeros. The sequencing data in the third data set hPSC has almost no zeros caused by data loss, which can better reflect the judgment ability of the five algorithms. As can be seen from **Figure 8**, in the Trapnell and Romanov data sets, the clustering effects of the five algorithms after missing value filling are not significantly different. After filled by scImpute, DrImpute, AdImpute, AutoImpute, and DeepImpute, the clustering results are improved compared with the raw data. However, from the experimental results of hPSC data set, we can see that the effect of AdImpute is significantly higher than the other four methods, which shows that AdImpute algorithm has good performance in identifying true zeros. In general, AdImpute performs best on the real data sets.

By comprehensively analyzing the results of the simulated data sets and the real data sets, we draw the following conclusions. scImpute prefers to regard the identified zeros as true zeros, so it performs well on the real data sets, but it does not perform well on the simulated data sets. DrImpute prefers to treat the identified zeros as the missing zeros, so it performs well in the simulated data sets, but it does not perform well in the real data sets. One of the limitations of DrImpute is that it considers only cell-level correlation using a simple hot deck approach. The performance of AutoImpute is not satisfactory on both the simulated data sets and the real data sets, but its effect on hPSC data set is better than that of scImpute, DrImpute, and DeepImpute. AutoImpute behaves ideally in retaining the most of true zeros present in the data. It is speculated that AutoImpute has a poor judgment ability to determine missing values, and most of the identified zeros are considered as true zeros. DeepImpute performs ordinarily on both the simulated data sets and the real data sets. It is designed for the bulk-RNAseq data and is suitable for handling large datasets. Its training and the prediction processes are separate, and DeepImpute tends to fail when the data show large heterogeneity and sparsity, which are two key characteristics of scRNA-seq data. AdImpute has minimally modified the expression of real biological silencing genes, and the filling effect is very robust.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Trapnell: Primary human myoblast scRNA-seq data, available in the GEO database, accession number GSE52529. hPSC: Human pluripotent stem cell scRNA-seq data, which can be obtained from the GEO database under the accession number GSE75748. Romanov: Mouse hypothalamus scRNA-seq data, which can be obtained in the GEO database under the accession number GSE74672.

AUTHOR CONTRIBUTIONS

JL provided the guidance during the whole research. YX and TX collected the data. LX, YX, TX, and XZ carried out the data analysis. YX and LX wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62172122, and the Fundamental Research Funds for the Central Universities, Jilin University under Grant No. 93K172021K04.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.739677/full#supplementary-material

REFERENCES

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 17:63. doi: 10.1186/s13059-016-0927-y

Chen, M., and Zhou, X. (2018). VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol.* 19:196.

Chu, L. F., Leng, N., Zhang, J., Hou, Z. G., Mamott, D., Vereide, D. T., et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17:173. doi: 10.1186/s13059-016-1033-x

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* 10:390. doi: 10.1038/s41467-018-07931-2

Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., and Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinform.* 19:220.

Hu, Z., Zu, S., and Liu, J. S. (2020). SIMPLEs: a single-cell RNA sequencing imputation strategy preserving gene modules and cell clusters variation. *NAR Genom. Bioinform.* 2:lqaa077.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., et al. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542. doi: 10.1038/s41592-018-0033-z

Kalisky, T., Oriel, S., Bar-Lev, T. H., Ben-Haim, N., Trink, A., Wineberg, Y., et al. (2018). A brief review of single-cell transcriptomic technologies. *Brief. Funct. Genomics.* 17, 64–76. doi: 10.1093/bfgp/elx019

Li, W. V., and Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* 9:997. doi: 10.1038/s41467-018-03405-7

Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNAseq data. *Genome Biol.* 18:59. doi: 10.1186/s13059-017-1188-0

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17:75. doi: 10.1186/s13059-016-0947-7

Peng, J. J., Guan, J. J., Hui, W. W., and Shang, X. Q. (2021a). A novel subnetwork representation learning method for uncovering disease-disease relationships. *Methods* 192, 77–84. doi: 10.1016/j.ymeth.2020.09.002

Peng, J. J., Xue, H. S., Wei, Z. Y., Tuncali, I., Hao, J. Y., and Shang, X. Q. (2021b). Integrating multi-network topology for gene function prediction using deep neural networks. *Brief. Bioinform.* 22, 2096–2105. doi: 10.1093/bib/bbaa 036

Qi, J., Zhou, Y., Zhao, Z., and Jin, S. (2021). SDImpute: a statistical block imputation method based on cell-level and gene-level information for dropouts in single-cell RNA-seq data. *PLoS Comput. Biol.* 17:e1009118.

Romanov, R. A., Zeisel, A., Bakker, J., Girach, F., Hellysaz, A., Tomer, R., et al. (2017). Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* 20, 176–188. doi: 10.1038/nn. 4462

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi: 10.1038/nrg3833

Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387.

Talwar, D., Mongia, A., Sengupta, D., and Majumdar, A. (2018). AutoImpute: autoencoder based imputation of single-cell RNA-seq Data. *Sci. Rep.* 8:16329. doi: 10.1038/s41598-018-34688-x

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudo temporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386.

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* 11:e1004333. doi: 10.1371/journal.pcbi.1004333

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* 174, 716–729.e27.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34, 1145–1160. doi: 10.1038/nbt.3711

Zhang, L., and Zhang, S. (2020). Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE ACM Trans. Comput. Biol. Bioinform.* 17, 376–389.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell.* 65, 631–643.e4.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 September 2021
doi: 10.3389/fgene.2021.756471

Landscape of T Cells Transcriptional and Metabolic Modules During HIV Infection Based on Weighted Gene Co-expression Network Analysis

Jianting Xu[1], *Jiahui Pan*[2], *Xin Liu*[1], *Nan Zhang*[3], *Xinyue Zhang*[2], *Guoqing Wang*[2]* and *Wenyan Zhang*[1]*

[1]Institute of Virology and AIDS Research, The First Hospital of Jilin University, Changchun, China, [2]College of Basic Medicine, Jilin University, Changchun, China, [3]College of Mathematics, Jilin University, Changchun, China

OPEN ACCESS

Edited by:
Jiajie Peng,
Northwestern Polytechnical University, China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Zhai Aixia,
Sun Yat-sen University, China
Qin Ma,
The Ohio State University, United States

***Correspondence:**
Guoqing Wang
qing@jlu.edu.cn
Wenyan Zhang
zhangwenyan@jlu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 10 August 2021
Accepted: 06 September 2021
Published: 16 September 2021

Citation:
Xu J, Pan J, Liu X, Zhang N, Zhang X, Wang G and Zhang W (2021) Landscape of T Cells Transcriptional and Metabolic Modules During HIV Infection Based on Weighted Gene Co-expression Network Analysis. Front. Genet. 12:756471. doi: 10.3389/fgene.2021.756471

Human immunodeficiency virus (HIV) causes acquired immunodeficiency syndrome (AIDS). HIV infection affects the functions and metabolism of T cells, which may determine the fate of patients; however, the specific pathways activated in different T-cell subtypes ($CD4^+$ and $CD8^+$ T cells) at different stages of infection remain unclear. We obtained transcriptome data of five individuals each with early HIV infection, chronic progressive HIV infection, and no HIV infection. Weighted gene co-expression network analysis was used to evaluate changes in gene expression to determine the antiviral response. An advanced metabolic algorithm was then applied to compare the alterations in metabolic pathways in the two T-cell subtypes at different infection stages. We identified 23 and 20 co-expressed gene modules in $CD4^+$ T and $CD8^+$ T cells, respectively. $CD4^+$ T cells from individuals in the early HIV infection stage were enriched in genes involved in metabolic and infection-related pathways, whereas $CD8^+$ T cells were enriched in genes involved in cell cycle and DNA replication. Three key modules were identified in the network common to the two cell types: *NLRP1* modules, *RIPK1* modules, and *RIPK2* modules. The specific role of NLRP1 in the regulation of HIV infection in the human body remains to be determined. Metabolic functional analysis of the two cells showed that the significantly altered metabolic pathways after HIV infection were valine, leucine, and isoleucine degradation; beta-alanine metabolism; and PPAR signaling pathways. In summary, we found the core gene expression modules and different pathways activated in $CD4^+$ and $CD8^+$ T cells, along with changes in their metabolic pathways during HIV infection progression. These findings can provide an overall resource for establishing biomarkers to facilitate early diagnosis and potential guidance for new targeted therapeutic strategies.

Keywords: HIV infection, T cell, transcriptional modules, metabolomics, weighted gene co-expression network analysis

INTRODUCTION

Acquired immunodeficiency syndrome (AIDS) is caused by human immunodeficiency virus (HIV) infection (Sepkowitz, 2001), a lentivirus belonging to the subgroup of retro-RNA viruses. HIV infection induces changes in T lymphocyte functions, leading to alterations in the entire immune system of the host and disruption of homeostasis. These hallmarks of HIV infection manifest differently based on the infection period (Weiss, 1993).

As a component of the host's immune defence system, T cells participate in a series of immune responses against HIV infection (Gupta and Saxena, 2021). Both $CD4^+$ and $CD8^+$ T cells participate in the host adaptive immune response against bacterial and viral infections. In particular, $CD4^+$ T cells can "help" the activity of other immune cells by releasing cytokines and small protein mediators, whereas $CD8^+$ T cells directly kill the target cells after activation in the human body (Hoyer et al., 2014). HIV mainly infects $CD4^+$ T lymphocytes. Clinically, HIV infection results in low blood $CD4^+$ T-cell levels. In addition, $CD4^+$ T cells directly inhibit HIV by promoting other T cells to resist viral infection (Johnson et al., 2015). $CD8^+$ T cells are widely distributed on the surface of inhibitory and cytotoxic T lymphocytes, and their kinetics differ from those of $CD4^+$ T cells during HIV infection (Xu et al., 2014). However, the overall molecular mechanisms underlying the changes and actions of $CD4^+$ and $CD8^+$ T cells after HIV infection remain to be elucidated.

Moreover, changes in metabolism also represent a key to understanding the immune response during pathogen invasion. Metabolism plays a fundamental role in supporting the growth, proliferation, and activation status of T cells (Palmer et al., 2016; Masson et al., 2017). For example, $CD8^+$ T cells increase oxidative phosphorylation and steadily increase the glycolysis rate, whereas $CD4^+$ T cells reduce fatty acid oxidation (Bantug et al., 2018). In HIV-1 infection, changes in cell metabolism affect the susceptibility of $CD4^+$ T cells; HIV-infected $CD4^+$ T cells exhibit elevated metabolic activity and metabolic potential compared with those of HIV-exposed but uninfected cells (Valle-Casuso et al., 2019). Detection of gene expression changes related to metabolism could provide insight into changes in metabolic pathway activity under HIV infection (Lee et al., 2012).

Weighted Correlation Network Analysis (WGCNA) is a method that can be used to analyze highly correlated gene modules in multiple samples and discover the relationship between the modules and specific functions (Langfelder and Horvath, 2012). It can provide panoramic information of T cell transcriptome modules after HIV infection. We performed WGCNA to assess changes in gene expression profiles in human $CD4^+$ T cells and $CD8^+$ T cells during HIV-1 infection. We then focused on acute HIV infection to explore the possible antiviral effects of the two cell types after their interaction with HIV-1 and to identify some key pathways and targets involved in the infection response. Ultimately, this study can highlight the metabolic changes occurring in T cells at different stages of HIV infection using the metabolic algorithm. These findings can show the panoramic transcription and metabolism modules and provide new insights for further biomarker discovery. This research can facilitate the early detection of HIV infection and ultimately the development of new strategies for effective infection control.

MATERIALS AND METHODS

Data Collection

We first searched the GEO database GSE6740 and downloaded the gene expression profiles from five individuals each with acute HIV infection, chronic progressive HIV infection, and no HIV infection. These data were reported and deposited to the GEO database by Hyrcza et al. (2007).

WGCNA and Module Recognition

WGCNA is a widely used data-mining method in genomic applications. We used the WGCNA software package in the R environment (R Foundation for Statistical Computing, Vienna, Austria) to construct a co-expression network of differentially expressed genes between HIV-infected and non-infected individuals. The algorithms were used to calculate the correlations between the levels of differentially expressed genes after the selection of an appropriate threshold (β), and then a scale-free network was constructed. We used the minimum value of β greater than 0.85 as the most suitable threshold, and then used the topological overlap matrix (TOM) (direct correlation + indirect correlation) between genes for hierarchical clustering to construct a clustering tree, which contained different gene modules represented by different colours. In this process of module identification, we set the minimum number of genes contained in the module to 50 (Zhang and Horvath, 2005).

Metabolic Pathway Activity Analysis

The number of metabolic genes enriched in a particular pathway was combined with the expressional values of the genes using the following formulas:

$$E_{ij} = \frac{\sum_{k=1}^{n_j} gik}{n_j} \tag{1}$$

where E_{ij} indicates the average expression level of the ith gene in the jth cell type, g_{ik} indicates the expression level of the ith gene in the kth sample, and n_j indicates the number of samples in the jth cell type;

$$r_{ij} = \frac{E_{ij}}{\frac{1}{N}\sum_{j}^{N} E_{ij}} \tag{2}$$

where N indicates the number of cell types, and r_{ij} represents the ratio of the average expression level of the ith gene in the jth cell type to the average level of the gene in all cell types. A ratio greater than 1 indicates that the gene expression level in the cell is higher than the average expression level in all cells, and a ratio below 1 indicates the opposite pattern; and

$$S_{tj} = \frac{\sum_{i=1}^{m_t} W_i \times r_{ij}}{\sum_{i=1}^{m_t} W_i} \tag{3}$$

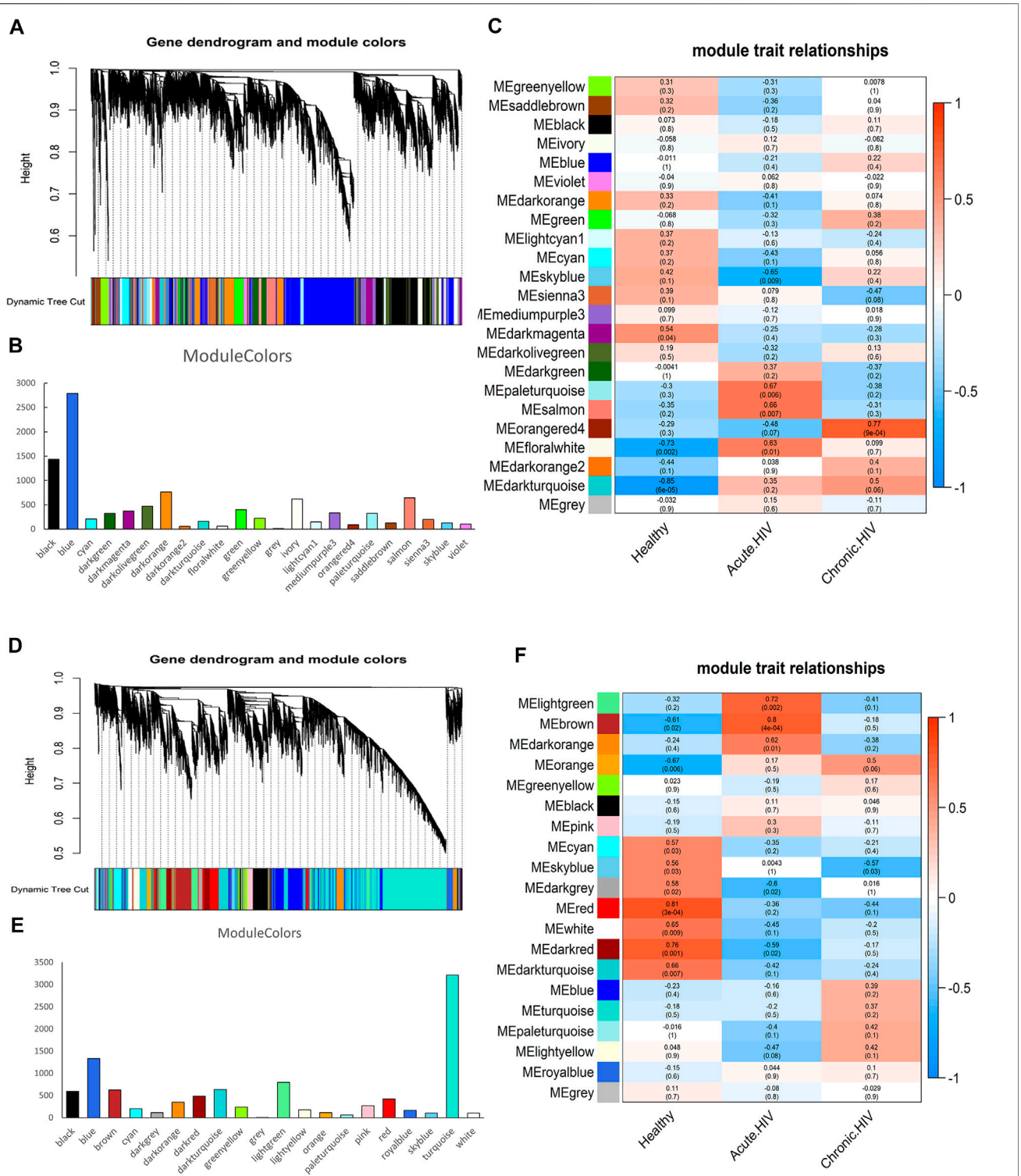

FIGURE 1 | Identification of the co-expressed gene modules. **(A)** Hierarchical clustering dendrogram of the co-expressed gene modules in $CD4^+$ T cells. **(B)** The number of genes contained in 23 modules of CD4 cells. **(C)** Correlation coefficients of module-trait relationships in $CD4^+$ T cells at different infection periods are indicated by different colours. **(D)** Hierarchical clustering dendrogram of the co-expressed gene modules in $CD8^+$ T cells. **(E)** The number of genes contained in 20 modules of CD8 cells. **(F)** Correlation coefficients of module-trait relationships in $CD8^+$ T cells at different infection periods are indicated by different colours.

where s_{tj} indicates the score value of the tth pathway in the jth cell type (the higher the score, the stronger the significance), m_t indicates the number of genes in the tth pathway, and W_i represents the weighted value of the ith gene (the reciprocal of the tth pathway metabolic gene) (Xiao et al., 2019).

Functional Annotation and Protein-Protein Interaction Network

Using clusterProfiler (an R package), we performed pathway enrichment analysis of the differentially expressed genes with respect to Gene Ontology terms (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, using default parameters. We then constructed a protein-protein interaction network based on these data using the STRING database in Cytoscape version 3.8.2. We also identified the chromosomal localisation of all genes in the target pathway by using the Ensembl and Genecards websites.

RESULTS

WGCNA Identification of Genetic Modules

In our study, the sample characteristics were divided into three stages as follows: Healthy, HIV acute infection, and HIV chronic infection. Then we choose the appropriate threshold (**Supplementary Figure S1**). We combined the relevant traits and modules of the sample for joint analysis to show the correlation between modules and traits using WGCNA's systems biology method. Different modules were represented by different colours. Each module contained a set of highly connected genes, and the genes in each module might participate in similar pathways or have the same biological functions. These modules ranged from large to small according to the number of genes that they contained.

The results of co-expression network analysis are shown in **Figure 1**. The number of genes in the module is shown in **Figures 1B,E** and **Supplementary Table S1**. The correlation coefficient of each module is shown in **Figures 1C,F**. We identified 23 co-expressed gene modules for $CD4^+$ T cells (**Figure 1A**). In terms of the number of genes contained in the module, MEblue was the largest module, containing 2,786 genes, whereas MEgrey was the smallest module, containing 13 genes. From the perspective of the correlation coefficient after infection, the module with the largest negative correlation coefficient was MEskyblue, with a value of −0.65; the module with the largest positive correlation coefficient was MEorangered4, with a value of 0.77. Each module has different functions. For example, Leukocyte transendothelial migration and FoxO signaling pathway are enriched in the MEskyblue.

We confirmed 20 co-expressed gene modules in $CD8^+$ T cells (**Figure 1D**). From the perspective of the number of genes contained in the module, MEturquoise was the largest module, containing 3,214 genes; MEgrey was the smallest module, containing only one gene. From the perspective of the correlation coefficient after infection, the module with the largest negative correlation coefficient was MEdarkgrey, with a value of −0.6; the module with the largest positive correlation coefficient was MEbrown, with a value of 0.8. In addition, the coefficient of the MElightgreen ranked second, reaching 0.72. RNA transport and Viral carcinogenesis are enriched in the MElightgreen.

Key Module-Activated Cell Processes

By calculating the correlation between the gene modules and the phenotype matrix, the key modules were screened out (i.e. those exhibiting higher correlations). The "salmon" module and the "orangered4" module were selected to represent the genes affected during early and chronic infection in $CD4^+$ T cells, respectively, whereas the "brown" module and the "skyblue" module were selected to represent early and chronic infection in $CD8^+$ T cells, respectively. We draw scatter plots of the relationship between gene saliency and module membership in the four modules respectively (**Figure 2**). Functional annotation of the key modules showed a distinct biological significance bias for each module (**Supplementary Table S2**) for example, the "salmon" module (early infection of $CD4^+$ T cells) was significantly enriched in inactivation of metabolic and infection-related pathways (**Figure 2A**), whereas the "orangered4" module (chronic infection of $CD4^+$ cells) was most significantly enriched in the TGF-beta signaling pathway and IL-17 signaling pathway, among others (**Figure 2B**). By contrast, the cell cycle and DNA replication were activated in the early infection $CD8^+$ T cells (**Figure 2C**), whereas the proteasome and sphingolipid metabolism were largely activated in $CD8^+$ T cells in the chronic infection stage (**Figure 2D**).

Critical Pathways During Acute HIV Infection

We focused on the similarities and differences in the activated gene pathways between $CD4^+$ and $CD8^+$ T cells during early infection. There were 20 special pathways among CD4 cells and 70 special pathways for CD8 cells. In total, 34 gene pathways (**Figure 3A**) overlapped between the two cell types. Among them, the top pathways mainly included the following: metabolic pathways, human immunodeficiency virus 1 infection, NOD-like receptor signaling pathway, and cAMP signaling pathway, among others. In addition to metabolic pathways and HIV infection pathways, two pathways with broad significance, NOD-like signaling pathways rank the top among other pathways. So we further focused on the NOD-like receptor (NLR) pathway. NLRs are type pattern recognition receptors for the host, which can recognise the pathogen-related molecular patterns of viruses to regulate antiviral innate immune signaling pathways, thereby regulating the innate antiviral immune response (Zheng, 2021). We first identified the chromosomal localisations for all of the genes enriched in the NLR pathway (**Figure 3B**), as well as the expression modules showing consistent direction of change (down- or up-regulated) in the two cells in the context of early and chronic infection (**Figures 3C–F**). The modules

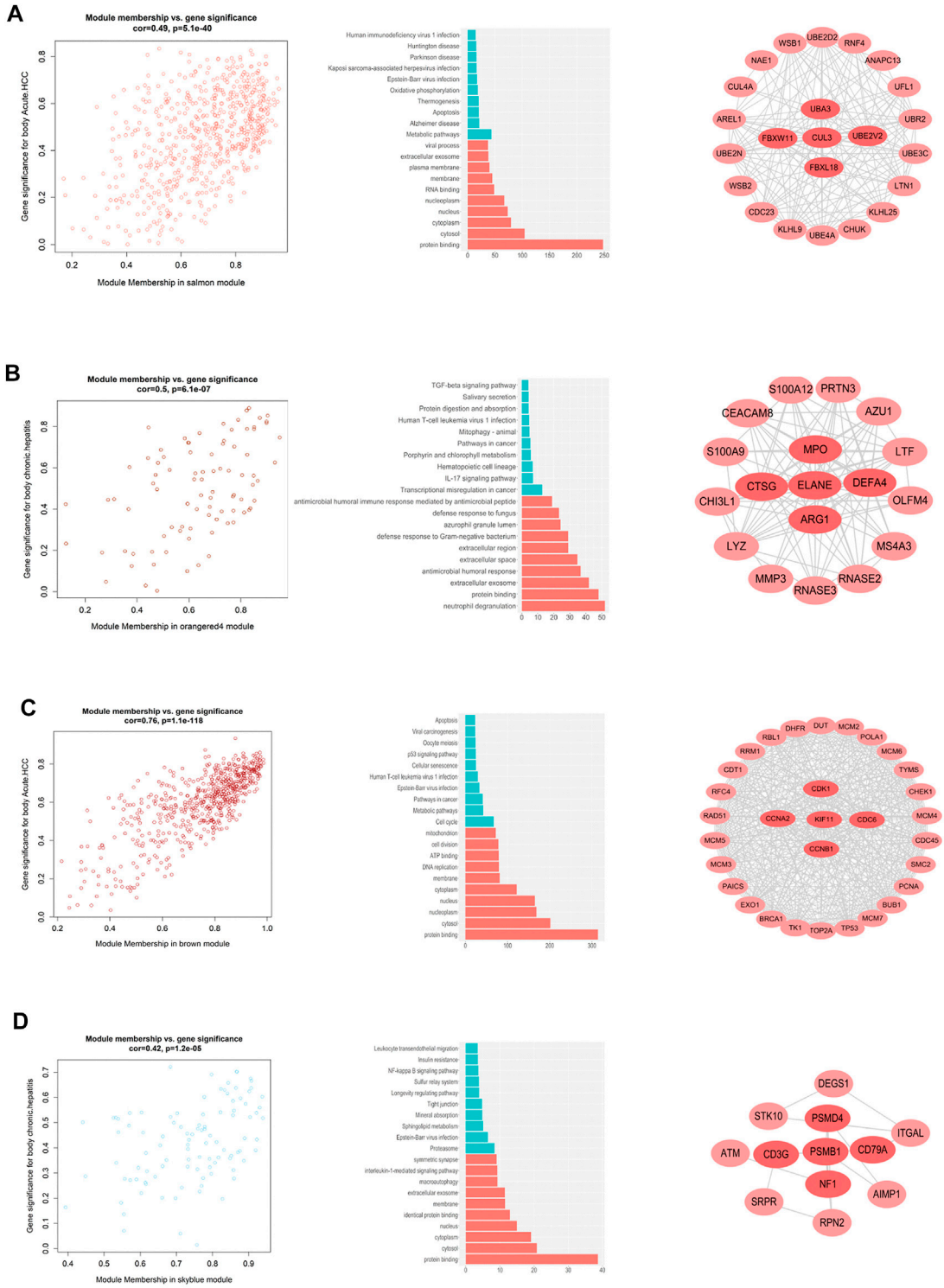

FIGURE 2 | Co-expressed gene modules that interact with other genes in different stages of HIV infection. **(A)** Distribution of genes in the salmon module, GO and KEGG pathway enrichment, and key genes in the early HIV infection stage of $CD4^+$ T cells. **(B)** Distribution of genes in the orange4 module, GO and KEGG pathway enrichment, and key genes of $CD4^+$ T cells in the chronic HIV infection stage. **(C)** Distribution of genes in the brown module, GO and KEGG pathway enrichment, and key genes in the early HIV infection of $CD8^+$ T cells. **(D)** Distribution of genes in the sky blue module, GO and KEGG pathway enrichment, and key genes in the chronic HIV infection of $CD8^+$ T cells.

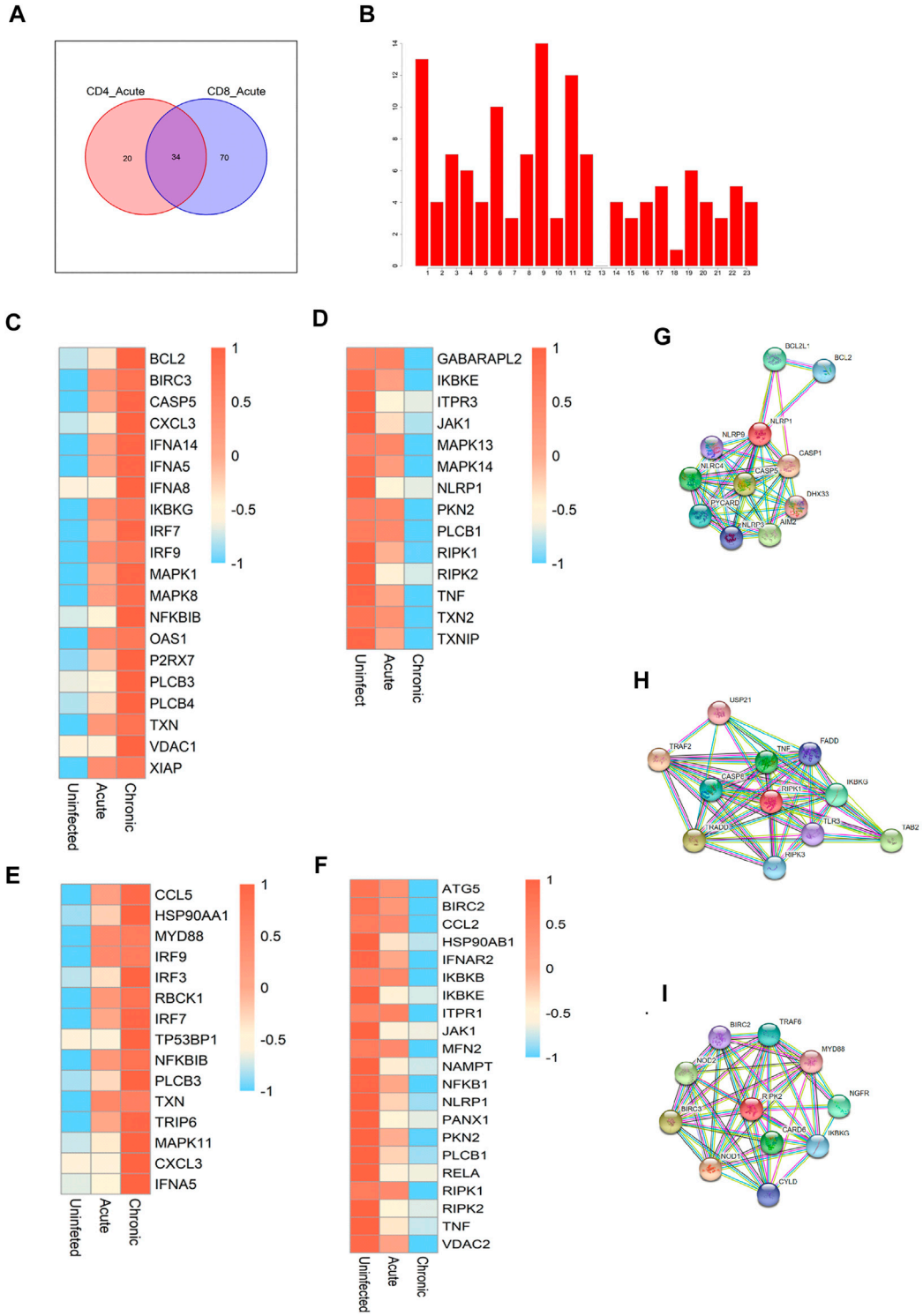

FIGURE 3 | Critical pathways during acute HIV infection. **(A)** Enrichment of pathways in $CD4^+$ and $CD8^+$ T cells during the early stage of HIV infection. The Venn diagram shows a total of 34 pathways common to the two cell types, with 20 $CD4^+$ T cell-specific pathways and 70 $CD8^+$ T cell-specific pathways. **(B)** Chromosomal localisation of all genes in the NOD-like receptor pathway. **(C,D)** Gene modules and co-upregulated gene modules in the NOD-like receptor pathway in $CD4^+$ T cells during different infection periods. **(E,F)** Co-upregulated and co-downregulated gene modules in $CD8^+$ T cells during different HIV infection periods. **(G–I)** Protein-protein interaction network centralised with respect to NLRP1 **(G)**, RIPL1 **(H)**, and RIPK2 **(I)**.

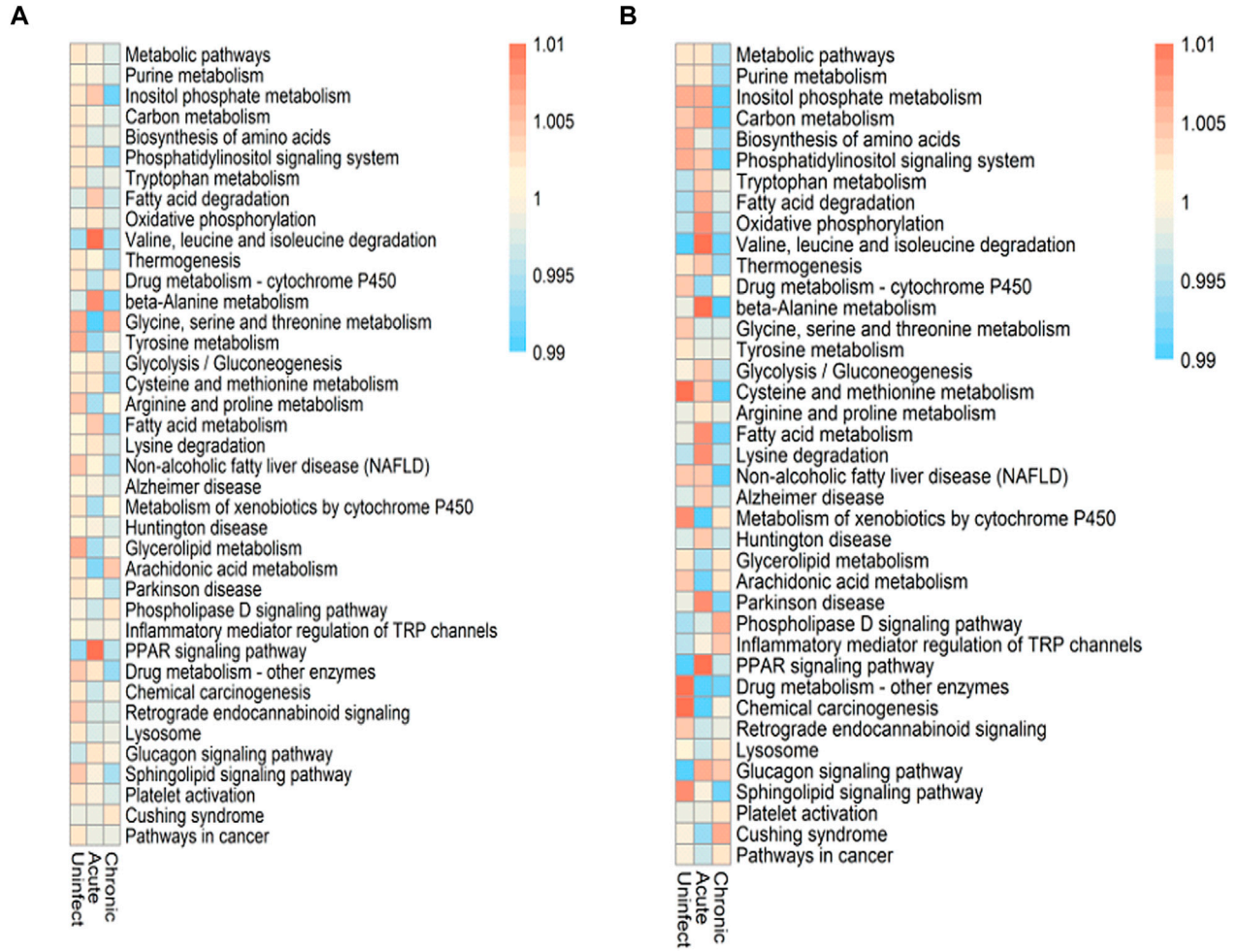

FIGURE 4 | Altered metabolic pathways during HIV infection. **(A)** Changes and scores of the metabolic pathways in $CD4^+$ T cells before and after infection and during different infection periods. **(B)** Changes and scores of the metabolic pathways of $CD8^+$ T cells before and after infection and during different infection periods.

that were co-up-regulated in CD4 and CD8 cells after infection included IRF9, IRF7, PLCB3, CXCL3, and TXN, among others. The co-down-regulated modules we identified included JAK1, PKN2, TNF, NLRP1, RIPK1, and RIPK2, among others. Among them, three core modules were selected, including NLR family pyrin domain containing 1 (NLRP1), receptor-interacting serine/threonine kinase 1 (RIPK1), and RIPK2. By constructing a protein–protein interaction network, we identified the genes that might interact with each other in the three modules (**Figures 3G–I**).

Differential activation of cell metabolism between $CD4^+$ and $CD8^+$ T cells before and after HIV infection.

There are a total of 9,700 genes in our data set, and 1,352 genes related to metabolism are collected. These metabolic genes come from metabolic pathways. We selected 296 genes expressed by both as the most basic metabolic genes in this study. Since $CD4^+$ and $CD8^+$ T cells play different roles in the immune response, we assessed the differential activation and characteristics of these two cell types and screened their co-expressed metabolic genes. The cell metabolism pathways were differentially activated before and after infection. HIV infection also appears to disrupt the metabolic balance between these two cell types; some metabolic pathways were activated, whereas others remained unchanged. As shown in **Figure 4**, the pathways that were significantly altered in both types of cells in early infection included valine, leucine, and isoleucine degradation; beta-alanine metabolism; and PPAR signaling pathways. These three pathways also showed the greatest change in $CD4^+$ T cells between the early infection and chronic infection stages. Among them, the PPAR signaling pathway also has significant changes in pathogen infections such as ZIKV and Neisseria meningitidis. More HIV-induced metabolic abnormalities were detected in $CD8^+$ T cells compared with those occurring in $CD4^+$ T cells. Before and after infection, the majority of pathways were changed in $CD8^+$ T cells, along with some pathways such as the arachidonic acid metabolism pathway that was unchanged. In addition to the three most significant pathways mentioned above, the oxidative

phosphorylation, fatty acid metabolism, and lysine degradation pathways also exhibited relatively large changes in $CD8^+$ T cells between the acute and chronic infection stages. Thus, changes in these metabolic pathways may be conducive for these two cell types to cope with HIV infection.

DISCUSSION

The primary function of $CD4^+$ T cells after HIV infection is related to DNA repair in response to DNA damage stimuli, along with positive regulation of cellular processes and other pathways, whereas $CD8^+$ T-cell functions after infection are mainly related to cell mitosis, signal transduction, and transmission (Xu et al., 2014). In addition, network-based methods have been widely used in biological data analysis (Peng et al., 2021a; Peng et al., 2021b). Therefore, we use WGCNA analysis to find that $CD4^+$ T cells from individuals in the early HIV infection stage were enriched in genes involved in metabolic and infection-related pathways, whereas $CD8^+$ T cells were enriched in genes involved in cell-related changes, including the cell cycle and DNA replication. During chronic HIV infection, CD4 cells are mainly enriched in pathways related to immune defense, such as IL-17 signaling pathway. CD8 cells are mainly enriched in proteasome and sphingolipid metabolism. This finding identified many other pathways altered in the two T-cell subtypes at different stages of HIV infection. It also expands evidence for the field to enrich overall understanding of HIV infection-related gene alterations and modules. However, in chronic infection, the two types of cells share fewer pathways. Therefore, when screening critical pathways, we choose acute infection, which makes it easier to find infection markers and therapeutic targets.

In addition, few studies have focused on the systematic characteristics of the two cells at different stages of HIV infection or the similarities and differences between the two cells at the same stage. We used GO annotations and KEGG pathways to analyze the core pathways in the 2 T cell types during early and chronic HIV infection, and then explored key co-expression modules among them. We identified three key down-regulation modules: NLRP1 module, RIPK1 module and RIPK2 module. The central gene of the module represents the function of the module to a certain extent. Among them, RIPK1 and RIPK2 are the key mediators of cell apoptosis and death, as well as the inflammatory pathways (Festjens et al., 2007). RIPK1 and RIPK2 can be cleaved by HIV-1 protease, which affects important biological processes in the body such as host defence pathways and cell death (Wagner et al., 2015). However, the specific role of NLRP1 in the regulation of HIV infection in the human body remains to be determined. NLR is a type of germline-encoded pattern recognition receptor, which is mainly involved in the cytosolic sensing mechanism to detect viral infections in the body. NLRs participate in immune signaling pathways, including inflammasomes, nuclear factor -kappa B, and type I interferon signaling (Lupfer and Kanneganti, 2013). In terms of viral infection, NLRs play an impor tant role in both innate and adaptive immunity. NLRP1 was the first protein identified to form an inflammasome and is a sensor for a variety of pathogens, which can activate an antibacterial or antiviral immune response (Chavarria-Smith and Vance, 2015; Chavarria-Smith et al., 2016). RIPK family members have also been documented to be related to NLRs. The association of RIPK and NLRP1 in this study further confirms their role in HIV infection, although further experimental studies are needed to explore their actual link. Moreover, the core genes identified in each module, and the specific types of genes in the modules corresponding to different infection stages and cell types could guide new therapeutic targets of HIV infection.

Cellular immune metabolism has become one of the hottest research topics in immunology (Medzhitov, 2015). Previous studies also showed that HIV infection led to upregulation of amino acid metabolism, the tricarboxylic acid cycle, and fatty acid metabolism in human $CD4^+$ T cells (Chan et al., 2007; Ringrose et al., 2008; Zhang et al., 2017). In our study, we utilised a novel algorithm to analyse differences in the metabolic pathways of $CD4^+$ and $CD8^+$ T cells before and after HIV infection. Our data demonstrate significant changes in three pathways of oxidative phosphorylation, fatty acid metabolism, and lysine degradation in CD8+T cells after early HIV infection compared with those assessed from individuals with chronic infection. The degree of metabolism of CD8 cells after infection is much stronger than that of CD4 cells. We have enriched the metabolic pathways of the two cells that are significantly altered in the early stage of HIV infection. These metabolic characteristics may be of great significance and warrant further investigation into identifying the mechanism of action of these two immune cell types after HIV infection.

In this study, we used WGCNA technology and metabolic algorithms to show a panoramic view of the core modules and metabolic pathways associated with HIV infection, providing new ideas and strategies for the development of HIV therapeutic targets and early diagnosis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GEO database—GSE6740.

AUTHOR CONTRIBUTIONS

JX wrote the manuscript. JP collected the data. NZ analyzed the data. XZ and XL did literature search. GW and WZ designed this study. The corresponding author have final responsibility for the decision to submit for publication. All authors approved the version of this manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (#81871699, #81930062, and #82072330); The epidemiology, early warning and response techniques of major infectious diseases in the Belt and Road Initiative (#2018ZX10101002).

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.cn) and Medjaden Bioscience Limited (Hong Kong, China) for editing and proofreading this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.756471/full#supplementary-material

Supplementary Figure 1 | Threshold filter. **(A)** The threshold of the module in CD4 cells is set to 9. **(B)** The screening of the module threshold in CD8 cells is set at 5.

Supplementary Table 1 | Each module of the two types of cells contains the number of genes.

Supplementary Table 2 | Key module functions. Two types of cell acute and chronic infections each have a key module, including KEGG enrichment and GO annotation.

REFERENCES

Bantug, G. R., Galluzzi, L., Kroemer, G., and Hess, C. (2018). The Spectrum of T Cell Metabolism in Health and Disease. *Nat. Rev. Immunol.* 18 (1), 19–34. doi:10.1038/nri.2017.99

Chan, E. Y., Qian, W.-J., Diamond, D. L., Liu, T., Gritsenko, M. A., Monroe, M. E., et al. (2007). Quantitative Analysis of Human Immunodeficiency Virus Type 1-Infected CD4 + Cell Proteome: Dysregulated Cell Cycle Progression and Nuclear Transport Coincide with Robust Virus Production. *J. Virol.* 81 (14), 7571–7583. doi:10.1128/JVI.00288-07

Chavarría-Smith, J., Mitchell, P. S., Ho, A. M., Daugherty, M. D., and Vance, R. E. (2016). Functional and Evolutionary Analyses Identify Proteolysis as a General Mechanism for NLRP1 Inflammasome Activation. *Plos Pathog.* 12 (12), e1006052. doi:10.1371/journal.ppat.1006052

Chavarría-Smith, J., and Vance, R. E. (2015). The NLRP1 Inflammasomes. *Immunol. Rev.* 265 (1), 22–34. doi:10.1111/imr.12283

Festjens, N., Vanden Berghe, T., Cornelis, S., and Vandenabeele, P. (2007). RIP1, a kinase on the crossroads of a cell's decision to live or die. *Cell Death Differ* 14 (3), 400–410. doi:10.1038/sj.cdd.4402085

Gupta, P. K., and Saxena, A. (2021). HIV/AIDS: Current Updates on the Disease, Treatment and Prevention. *Proc. Natl. Acad. Sci. India, Sect. B Biol. Sci.*, 1–16. doi:10.1007/s40011-021-01237-y

Hoyer, S., Prommersberger, S., Pfeiffer, I. A., Schuler-Thurner, B., Schuler, G., Dörrie, J., et al. (2014). Concurrent Interaction of DCs with CD4+and CD8+T Cells Improves Secondary CTL Expansion: It Takes Three to Tango. *Eur. J. Immunol.* 44 (12), 3543–3559. doi:10.1002/eji.201444477

Hyrcza, M. D., Kovacs, C., Loutfy, M., Halpenny, R., Heisler, L., Yang, S., et al. (2007). Distinct Transcriptional Profiles in *Ex Vivo* CD4 + and CD8 + T Cells Are Established Early in Human Immunodeficiency Virus Type 1 Infection and Are Characterized by a Chronic Interferon Response as Well as Extensive Transcriptional Changes in CD8 + T Cells. *J. Virol.* 81 (7), 3477–3486. doi:10.1128/JVI.01552-06

Johnson, S., Eller, M., Teigler, J. E., Maloveste, S. M., Schultz, B. T., Soghoian, D. Z., et al. (2015). Cooperativity of HIV-specific Cytolytic CD4 T Cells and CD8 T Cells in Control of HIV Viremia. *J. Virol.* 89 (15), 7494–7505. doi:10.1128/JVI.00438-15

Langfelder, P., and Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. *J. Stat. Softw.* 46 (11). doi:10.18637/jss.v046.i11

Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., et al. (2012). Improving Metabolic Flux Predictions Using Absolute Gene Expression Data. *BMC Syst. Biol.* 6, 73. doi:10.1186/1752-0509-6-73

Lupfer, C., and Kanneganti, T.-D. (2013). The Expanding Role of NLRs in Antiviral Immunity. *Immunol. Rev.* 255 (1), 13–24. doi:10.1111/imr.12089

Masson, J. J. R., Murphy, A. J., Lee, M. K. S., Ostrowski, M., Crowe, S. M., and Palmer, C. S. (2017). Assessment of Metabolic and Mitochondrial Dynamics in CD4+ and CD8+ T Cells in Virologically Suppressed HIV-Positive Individuals on Combination Antiretroviral Therapy. *PLoS One* 12 (8), e0183931. doi:10.1371/journal.pone.0183931

Medzhitov, R. (2015). Bringing Warburg to Lymphocytes. *Nat. Rev. Immunol.* 15 (10), 598. doi:10.1038/nri3918

Palmer, C. S., Henstridge, D. C., Yu, D., Singh, A., Balderson, B., Duette, G., et al. (2016). Emerging Role and Characterization of Immunometabolism: Relevance to HIV Pathogenesis, Serious Non-AIDS Events, and a Cure. *J.I.* 196 (11), 4437–4444. doi:10.4049/jimmunol.1600120

Peng, J., Guan, J., Hui, W., and Shang, X. (2021a). A Novel Subnetwork Representation Learning Method for Uncovering Disease-Disease Relationships. *Methods* 192, 77–84. doi:10.1016/j.ymeth.2020.09.002

Peng, J., Xue, H., Wei, Z., Tuncali, I., Hao, J., and Shang, X. (2021b). Integrating Multi-Network Topology for Gene Function Prediction Using Deep Neural Networks. *Brief. Bioinform.* 22 (2), 2096–2105. doi:10.1093/bib/bbaa036

Ringrose, J. H., Jeeninga, R. E., Berkhout, B., and Speijer, D. (2008). Proteomic Studies Reveal Coordinated Changes in T-Cell Expression Patterns upon Infection with Human Immunodeficiency Virus Type 1. *J. Virol.* 82 (9), 4320–4330. doi:10.1128/JVI.01819-07

Sepkowitz, K. A. (2001). AIDS - the First 20 Years. *N. Engl. J. Med.* 344 (23), 1764–1772. doi:10.1056/NEJM200106073442306

Valle-Casuso, J. C., Angin, M., Volant, S., Passaes, C., Monceaux, V., Mikhailova, A., et al. (2019). Cellular Metabolism Is a Major Determinant of HIV-1 Reservoir Seeding in CD4+ T Cells and Offers an Opportunity to Tackle Infection. *Cel Metab.* 29 (3), 611–626 e615. doi:10.1016/j.cmet.2018.11.015

Wagner, R. N., Reed, J. C., and Chanda, S. K. (2015). HIV-1 Protease Cleaves the Serine-Threonine Kinases RIPK1 and RIPK2. *Retrovirology* 12, 74. doi:10.1186/s12977-015-0200-6

Weiss, R. (1993). How Does HIV Cause AIDS? *Science* 260 (5112), 1273–1279. doi:10.1126/science.8493571

Xiao, Z., Dai, Z., and Locasale, J. W. (2019). Metabolic Landscape of the Tumor Microenvironment at Single Cell Resolution. *Nat. Commun.* 10 (1), 3763. doi:10.1038/s41467-019-11738-0

Xu, C., Ye, B., Han, Z., Huang, M., and Zhu, Y. (2014). Comparison of Transcriptional Profiles between CD4+ and CD8+ T Cells in HIV Type 1-infected Patients. *AIDS Res. Hum. Retroviruses* 30 (2), 134–141. doi:10.1089/AID.2013.0073

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. doi:10.2202/1544-6115.1128

Zhang, L.-L., Zhang, Z.-N., Wu, X., Jiang, Y.-J., Fu, Y.-J., and Shang, H. (2017). Transcriptomic Meta-Analysis Identifies Gene Expression Characteristics in Various Samples of HIV-Infected Patients with Nonprogressive Disease. *J. Transl Med.* 15 (1), 191. doi:10.1186/s12967-017-1294-5

Zheng, C. (2021). The Emerging Roles of NOD-like Receptors in Antiviral Innate Immune Signaling Pathways. *Int. J. Biol. Macromolecules* 169, 407–413. doi:10.1016/j.ijbiomac.2020.12.127

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 October 2021
doi: 10.3389/fgene.2021.763259

The Causal Effects of Insomnia on Bipolar Disorder, Depression, and Schizophrenia: A Two-Sample Mendelian Randomization Study

*Peng Huang[1†], Yixin Zou[1†], Xingyu Zhang[2], Xiangyu Ye[1], Yidi Wang[1], Rongbin Yu[1] and Sheng Yang[3]**

[1]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China, [2]Thomas E. Starzl Transplantation Institute, University of Pittsburgh Medical Center, University of Pittsburgh, Pittsburgh, PA, United States, [3]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:
Tao Wang,
Northwestern Polytechnical University, China

Reviewed by:
Ping Zeng,
Xuzhou Medical University, China
Wenlong Ren,
Nantong University, China
Yue Fan,
Xi'an Jiaotong University, China

***Correspondence:**
Sheng Yang
yangsheng@njmu.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 23 August 2021
Accepted: 13 September 2021
Published: 11 October 2021

Citation:
Huang P, Zou Y, Zhang X, Ye X, Wang Y, Yu R and Yang S (2021) The Causal Effects of Insomnia on Bipolar Disorder, Depression, and Schizophrenia: A Two-Sample Mendelian Randomization Study. Front. Genet. 12:763259. doi: 10.3389/fgene.2021.763259

Psychiatric disorder, including bipolar disorder (BD), major depression (MDD), and schizophrenia (SCZ), affects millions of persons around the world. Understanding the disease causal mechanism underlying the three diseases and identifying the modifiable risk factors for them hold the key for the development of effective preventative and treatment strategies. We used a two-sample Mendelian randomization method to assess the causal effect of insomnia on the risk of BD, MDD, and SCZ in a European population. We collected one dataset of insomnia, three of BD, one of MDD, and three of SCZ and performed a meta-analysis for each trait, further verifying the analysis through extensive complementarity and sensitivity analysis. Among the three psychiatric disorders, we found that only insomnia is causally associated with MDD and that higher insomnia increases the risk of MDD. Specifically, the odds ratio of MDD increase of insomnia is estimated to be 1.408 [95% confidence interval (CI): 1.210–1.640, p = 1.03E-05] in the European population. The identified causal relationship between insomnia and MDD is robust with respect to the choice of statistical methods and is validated through extensive sensitivity analyses that guard against various model assumption violations. Our results provide new evidence to support the causal effect of insomnia on MDD and pave ways for reducing the psychiatric disorder burden.

Keywords: insomnia, bipolar disorder, depression, schizophrenia, two-sample mendelian randomization, genome-wide association study

INTRODUCTION

Insomnia disorder is predominantly characterized by dissatisfaction with sleep duration or quality and difficulties in initiating or maintaining sleep (Morin et al., 2015; Winkelman, 2015). Most cross-sectional and longitudinal studies have also shown that insomnia increases the risks of acute myocardial infarction and coronary heart disease, heart failure, hypertension, diabetes, and death, particularly when insomnia is accompanied by a short total sleep duration (<6 h per night) (Chen et al., 2013; Morin et al., 2015; Parthasarathy et al., 2015; Grandner et al., 2016; Javaheri and Redline, 2017; Bertisch et al., 2018; Dong and Yang, 2019). Emerging evidence show that insomnia associates to both incident and some recurrent psychiatric disorders, including major depression disorder

(MDD), anxiety disorder, substance use problems, and suicidality. In addition, a wide range of sociodemographic correlates of insomnia have been identified and include advanced age, female sex, low socioeconomic status, unemployment, and psychological distress. Although insomnia results from environmental factors, it is, in part, attributable to genetic factors (Winkelman, 2015).

The generation and development of psychiatric disorders are influenced by genetic and environmental factors (Sklar et al., 2011; Nagel et al., 2018; Ruderfer et al., 2018; Peng et al., 2020; Peng et al., 2021a; Peng et al., 2021b). For genetic factors, based on genome-wide association analysis (GWAS), Purcell *et al.* implicate the major histocompatibility complex, constructed a polygenic risk score (PRS) of schizophrenia (SCZ) and verified that the PRS also predicted bipolar disorder (BD) (Purcell et al., 2009). For environmental factors, using a case–control study, Palagini *et al.* found that insomnia played a mediating role between early life stress and the clinical manifestations of BD, and assessing the evolution of insomnia symptoms can provide a basis for the characteristics and treatment strategies of BD (Palagini et al., 2021). In addition, the result of longitudinal epidemiological studies shows that sleep disturbances and insomnia increase the risk of MDD after 1–3 years (Riemann and Voderholzer, 2003; Franzen and Buysse, 2008). Studies have also shown that up to 80% of patients with SCZ report symptoms of insomnia (Stummer et al., 2018). Sleep disorders have been shown to increase the risk of cognitive impairment and recurrence in patients with schizophrenia (Stummer et al., 2018). However, all these findings are summarized from either observational studies or pilot randomized controlled trials and prone to selection bias, especially unobserved confounding factors—that is, correlation cannot be simply equal to causal association. It is essential and urgent to further investigate the causal association between insomnia and psychiatric disorders, including BD, MDD, and SCZ.

Based on Mendel's law of inheritance—that is, parental alleles are randomly assigned to offspring—Mendelian randomization (MR), an advanced statistical method, treats single-nucleotide polymorphism (SNP) as an instrumental variable (IV) to adjust the effect of confounders and identifies the causal relationship between two traits (Davey Smith and Hemani, 2014; Paternoster et al., 2017). Then, when we regarded SNPs both with association to insomnia and without association to psychiatric disorders as IVs, MR can establish the causal relationship between insomnia and psychiatric disorders. Because genetic variants are fixed at conception and cannot be modified subsequently, MR can overcome a possible reverse causation. MR assumes that if insomnia causes psychiatric disorders, SNP related to insomnia causes psychiatric disorders through the insomnia pathway. Emerging large-scale GWAS of insomnia and psychiatric disorders gives us opportunities to use MR to study the causal relationship between them (Sleiman and Grant, 2010).

In the present study, our main aim is to investigate the causal relationship between insomnia and three psychiatric disorders (BD, MDD, and SCZ) in a European ancestry. To achieve the aim, we used the summary statistics of eight datasets (including 386,533 samples of insomnia and 719,027 samples of three psychiatric disorders) to perform a series of two-sample MR to comprehensively elucidate the potential causal association between insomnia and BD, MDD, and SCZ. In addition, to ensure the validity of the results of MR, we performed three sensitivity analyses, including heterogeneity test, pleiotropy test, and leave-one-out (LOO) test, and reverse-direction MR analyses (Zeng et al., 2019; Zeng and Zhou, 2019; Gormley et al., 2021).

TABLE 1 | Summary of the meta-datasets for four traits.

Trait	N_{SNP}	N_{sample}	Prev	h^2_O	h^2_L	λ_{GC}	Intcp
Insomnia	7,213,582	386,533	0.283	0.046	0.082	1.310	1.015
BD	9,018,454	78,638	0.366	0.405	0.286	1.421	1.080
MDD	7,743,682	500,199	0.341	0.060	0.067	1.453	1.00
SCZ	8,679,614	140,190	0.399	0.295	0.170	1.637	1.044

MATERIALS AND METHODS

GWAS Meta-Analysis

We collected eight datasets of insomnia and three psychiatric disorders from the GWAS-ATLAS (https://atlas.ctglab.nl/) (Tian et al., 2020), including one insomnia dataset (Jansen et al., 2019), three BD datasets (Smith et al., 2009; Hou et al., 2016; Ruderfer et al., 2018), one MDD dataset (Howard et al., 2019), and three SCZ datasets (Manolio et al., 2007; Ripke et al., 2013; Pardiñas et al., 2018). The insomnia summary statistics was estimated from the UK Biobank datasets with 386,533 individuals (Prev. = 0.283). The MDD summary statistics was estimated from the UK Biobank datasets with 500,199 individuals (Prev. = 0.341). The three BD summary statistics had 34,950 individuals (Prev. = 0.219), 2,035 individuals (Prev. = 0.492), and 41,653 individuals (Prev. = 0.483), repectively. The three SCZ summary statistics had 32,143 individuals (Prev. = 0.430), 2,729 individuals (Prev. = 0.495), and 105,318 individuals (Prev. = 0.386), repectively. The three studies for BD and MDD were without any overlap individuals. All summary statistics were estimated in the European ancestry. Then, we filtered out SNPs 1) with INFO <0.6, 2) with MAF <0.01, 3) with palindromic allele, and 4) whose OR was larger or smaller than mean ±3 SD. Finally, we obtained 7,213,582, 9,018,454, 7,743,682, and 8,679,614 SNPs for the four traits. Details of the meta-dataset and the three datasets for BD and SCZ are shown in **Table 1** and **Supplementary Table S1**.

Furthermore, to obtain an accurate and robust estimation for each variant, we performed GWAS meta-analysis for each trait using METAL (v2011-03-25) (Willer et al., 2010). To control the population stratification, we set the option *GENOMICCONTROL* to on. In addition, we used Linkage Disequilibrium SCore regression (LDSC) (v1.0.1) to estimate both the observed and liability observed heritability (h^2) for each trait. We set the population prevalence (*--pop-prev*) for the four traits to 0.300, 0.020, 0.086, and 0.010 to estimate liability heritability, respectively (Ayuso-Mateos et al., 2001; Roth, 2007; Di Luca et al., 2011). We also estimated the genetic correlation (R_g) between them in the GWAS analysis results (Tylee et al., 2018).

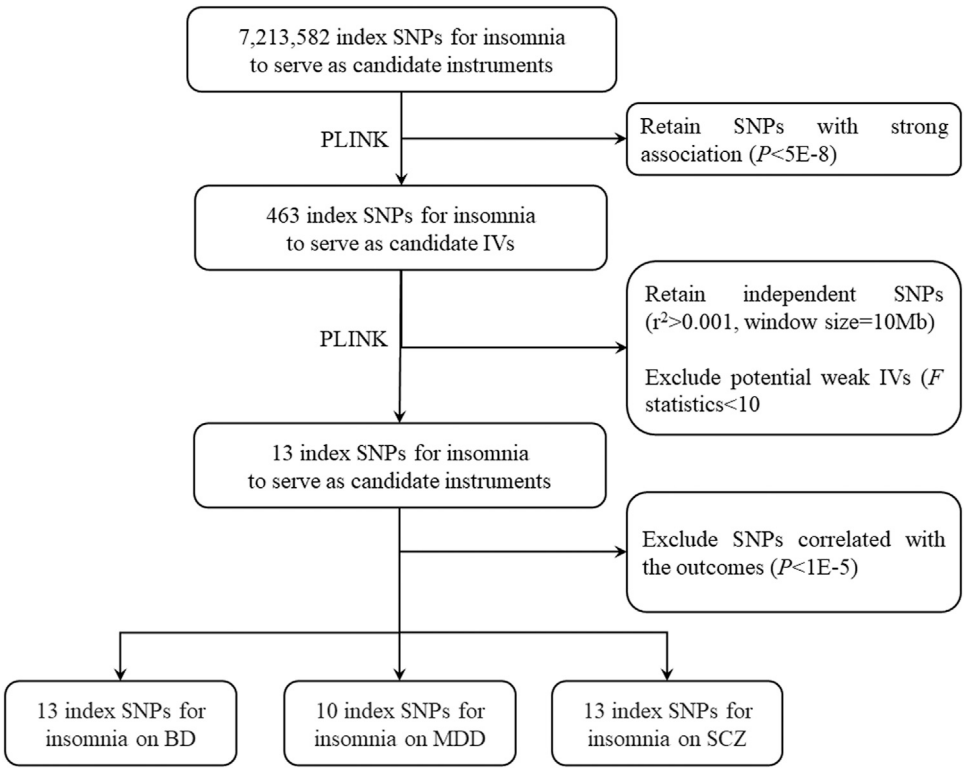

FIGURE 1 | Flow chart for instrumental variable (IV) selection. The flow chart shows the selection process of insomnia IVs to estimate the causal effects on bipolar disorder (BD), major depression (MDD), and schizophrenia (SCZ). First, we use $p < 5.00E-8$ to select index single-nucleotide polymorphisms (SNPs) to ensure that they strongly associate with insomnia. Second, we use $r^2 > 0.001$ in the range of 10,000 Mb to select independent index SNPs. We treat the EUR of 1000 Genome Project as the reference panel. The first two steps are completed by PLINK. Finally, we obtain 13 IVs on BD, 10 IVs on MDD, and 13 IVs on SCZ.

IV Selection

Based on the meta-datasets, we followed the strict selection procedure for selecting IVs in other previous MR studies (Zeng et al., 2019; Dong et al., 2021) (**Figure 1**). First, we retained 463 variants for insomnia with a P-value smaller than 5.00E-8. Second, we excluded 450 highly correlated variants with r^2 greater than 0.001 in the range of 10 Mb. In addition, following Zeng et al. (2019), we used *F* statistic to test for weak IVs, and no variant was excluded with a minimum *F* statistic of 39.37. Finally, we retained a total of 13 independent candidate IVs for studying the causal relationship between insomnia and BD, MDD, and SCZ. The details are shown in **Supplementary Table S2**.

We performed three two-sample MR analyses, including inverse variance weighted (IVW), MR-Egger, and weighted median (WM) method, to estimate the potential causal effect of insomnia to BD, MDD, and SCZ (Bowden et al., 2015; Bowden et al., 2016a). Without consideration for the intercept term, IVW regarded the reciprocal of the outcome variance as the weight. Because of no pleiotropy assumption, IVW was biased when pleiotropy exists (Bowden et al., 2015). Differently to IVW, MR-Egger used an intercept term to measure the horizontal pleiotropy between IVs (Bowden et al., 2016b). The weighted median method assumed that variables that account for at least 50% of the total IVs were valid, so the causal effects could be estimated consistently (Bowden et al., 2016a). We also used MR-Egger intercept to test pleiotropy (Hemani et al., 2018; Verbanck et al., 2018; Ong and MacGregor, 2019). All the analyses are performed by R software (v4.1.1). We specially used TwoSampleMR R package (v0.5.6) to perform a MR analysis.

Sensitivity Analysis

Following Noyce et al. (2017), Zeng et al. (2019), and Zeng and Zhou (2019), we performed a sensitivity analysis to evaluate the potential violations of the model assumptions in the MR analysis: 1) heterogeneity test, 2) pleiotropic test, and 3) LOO sensitivity test. First, heterogeneity analysis estimates heterogeneity between IVs. If heterogeneity existed, it is hardly to direct combinations of IVs. We used the *P*-value of *Q* statistics (P_Q) < 0.05 as the significant level. Second, we used MR-PRESSO to test pleiotropy, resulting in serious deviations in MR (Hemani et al., 2018; Ong and MacGregor, 2019). Finally, by gradually excluding each variant, LOO estimated the causal effect of the remaining variants and tested whether the difference between each causal effect is significant. Ideally, defining no significant difference meant a robust result (Noyce et al., 2017). The statistically significant level was set to 0.05.

Reverse-Direction MR Analyses

We also performed reverse-direction MR to assess the potential reverse causal effects of BD, MDD, and SCZ on insomnia.

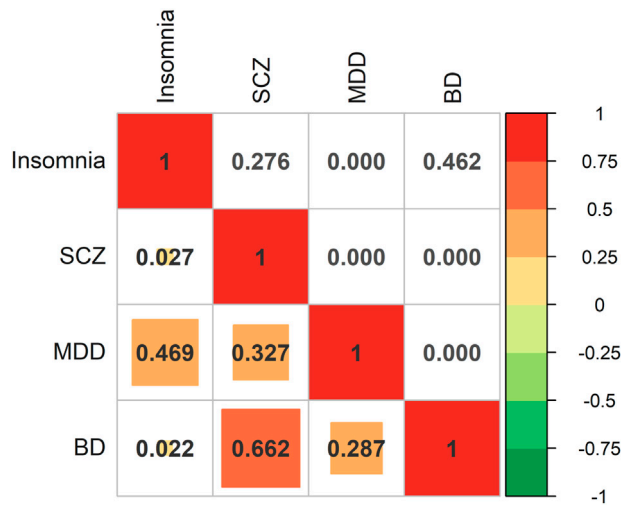

FIGURE 2 | The genetic correlation of four traits using Linkage Disequilibrium SCore regression. The heat map shows the R_g (lower triangle) and its P-value (upper triangle). The red color shows a more positive correlation, while the blue color shows a more negative correlation. We use abbreviations to indicate various psychiatric disorders: BD, bipolar disorder; MDD, major depression; SCZ, schizophrenia.

Following Savage et al. (2018) and Dong et al. (2021), we used the same settings as the abovementioned MR analysis (p = 5.00E-8, r^2 = 0.001, and window size = 10 Mb). We obtained 36 IVs for BD, 44 IVs for MDD, and 50 IVs for SCZ. We used these IVs of three psychiatric disorders to perform reverse causal inferences on insomnia to assess the potential reverse causal effects. The reverse-direction MR analysis process is the same as previously described.

RESULTS

Summary of GWAS Meta-Data and Genetic Correlation

We used the meta-analysis datasets to estimate the genetic correlation. The genetic inflation factor (λ_{gc}) of insomnia is 1.310 (LDSC intercept: 1.015), the λ_{gc} of BD is 1.421 (LDSC intercept: 1.080), the λ_{gc} of MDD is 1.453 (LDSC intercept: 1.000), and the λ_{gc} of SCZ is 1.637 (LDSC intercept: 1.044). The LDSC of the four traits are not larger than 1, which indicates that the meta-datasets are without population stratification. Using GWAS summary statistics to estimate SNP-based observed and liability heritability, these are 0.046 and 0.082 for insomnia, 0.405 and 0.286 for BD, 0.060 and 0.067 for MDD, and 0.295 and 0.170 for SCZ, respectively (**Table 1**). We use Manhattan plot and qqplot to show the GWAS results for the four traits (**Supplementary Figure S1**).

In addition, we assessed the genetic correlation between BD, MDD, SCZ, and insomnia using cross-trait LDSC. Insomnia was significantly genetically correlated to MDD (R_g = 0.469, p = 2.01E-70), while it was not significantly genetically correlated to BD (R_g = 0.022, p = 0.462) and SCZ (R_g = 0.027, p = 0.276). As expected, we defined three significant genetic correlations between the three psychiatric disorders: genetic correlation between BD and MDD (R_g = 0.287, p = 5.72E-26), between BD and SCZ (R_g = 0.662, p = 5.6E-283), and between MDD and SCZ (R_g = 0.327, p = 4.91E-42) (**Figure 2**).

MR Analysis

We use the 13 potential IVs of insomnia with the three psychiatric disorders one by one. Specifically, three psychiatric disorders had 13, 10, and 13 IVs, respectively (**Supplementary Table S2**). Based on different assumptions, we estimate the potential causal effect by all four models, including IVW (fixed- and random-effects model), MR-Egger, and WM. We use the forest plot to show the potential causal effect of the four methods, scatter plot to show the IV effect of insomnia and three psychiatric disorders, and funnel plot to show the relationship between effect of MR model and effect of each SNP (**Figure 3**, **Supplementary Figures S2, S3**, **Supplementary Tables S3–S5**).

For MDD, the estimated OR from fixed-effects IVW method is 1.288 (95% CI: 1.189–1.395), with p = 5.630E-11. As expected, the result of the random-effects IVW method (OR = 1.288, 95% CI: 1.091–1.520, p = 0.003) is similar to that of the random-effects IVW. However, the result of WM (OR = 1.076, 95% CI: 0.915–1.216, p = 0.374) and MR-Egger (OR = 0.916, 95% CI: 0.599–1.401, p = 0.696) is not similar to that of IVW (**Figure 4**, **Supplementary Table S4**). The abovementioned results indicate that the risk of MDD increases with the increasing level of insomnia. We should use the result of the sensitivity analysis to determine which one is the main result.

For BD, in terms of the fixed-effects IVW method, the estimated OR of insomnia is 1.216 (95% CI: 0.974–1.518, p = 0.084). As expected, the result of the random-effects IVW method is similar to that of the fixed-effects method, with OR = 1.216 (95% CI: 0.801–1.845) and p = 0.358. The results of WM (OR = 1.351, 95% CI: 0.948–1.917, p = 0.096) and MR-Egger (OR = 1.909, 95% CI: 0.506–7.197, p = 0.360) are similar (**Supplementary Figure S2**, **Supplementary Table S3**). Unfortunately, the results of all MR methods are not significant, suggesting that there might be no potential causal association for insomnia on BD. The specific results have to be verified after a sensitivity analysis.

Finally, for SCZ, the estimated OR of insomnia by the fixed-effects IVW method is 0.787 (95% CI: 0.630–0.983, p = 0.035), while the OR from the random-effect model is 0.787 (95% CI: 0.479–1.292, p = 0.344). In addition, the results of the weighted median method (OR = 0.604, 95% CI: 0.413–0.883, p = 0.009) and MR-Egger (OR = 0.566, 95% CI: 0.109–2.950, p = 0.513) are different (**Supplementary Figure S3**, **Supplementary Table S5**). Similarly, which specific result is representative also needs to be determined after the sensitivity analysis.

Sensitivity Analyses

Using three kinds of MR methods, we identify the potential causal relationship of insomnia on MDD (IVW method) and SCZ (only WM method). We performed a series of sensitivity analyses to assess whether the results obtained are robust, whether there is potential bias (such as pleiotropy and data heterogeneity), and

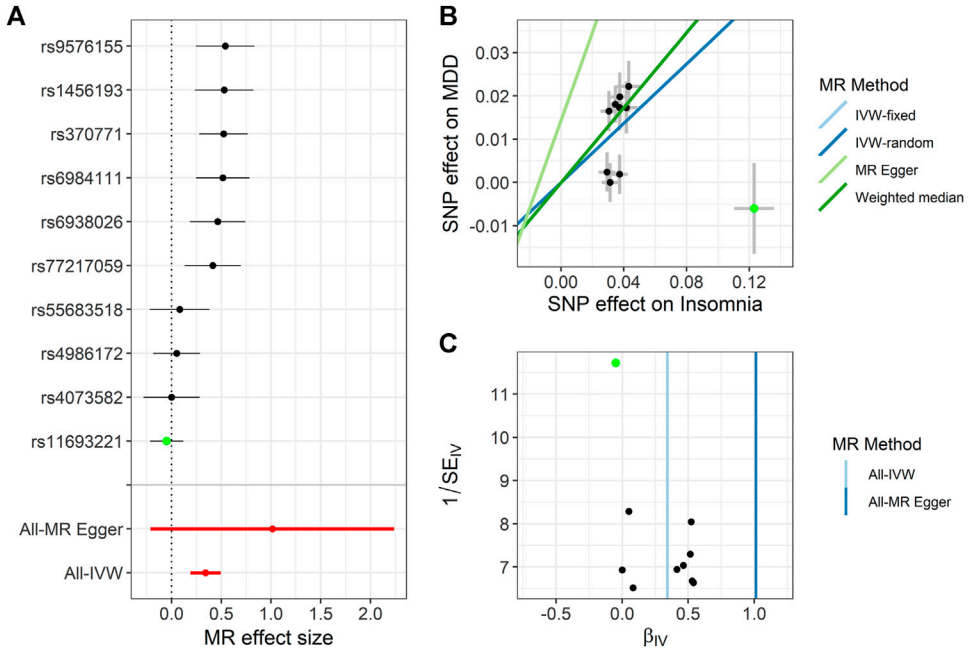

FIGURE 3 | Summary of the Mendelian randomization (MR) analysis for insomnia on major depression (MDD). **(A)** MR effect size of each instrumental variable (IV), MR-Egger, and inverse variance weighted (IVW). **(B)** Scatter plot of causal effects of insomnia on MDD. We use vertical and horizontal black lines to show 95% CI of the estimated effect of IVs on MDD (x-axis) and that on insomnia (y-axis), respectively. We use the blue line to show the IVW random-effects model. The potential SNP outlier (rs11693221) is highlighted in green. **(C)** Funnel plot of the causal effect of insomnia on MDD. Each point represents the estimated causal effect of each IV. The vertical dark blue line represents the causal effect estimate obtained using the MR-Egger method; the light blue line represents the causal effect estimate obtained using the IVW method. The potential outlier (rs11693221) is highlighted in green.

whether there is a certain IV that seriously affects the outcome variable.

First, we conducted a heterogeneity analysis. Based on IVW, the P_Q values of BD, MDD, and SCZ are 2.83E-5, 1.39E-5, and 2.4E-8, respectively. Following Zeng et al. (2019), we selected the result from the random-effects model or deleted SNPs with *P*-value < 1.00E-5. Because of the similarity between the fix- and random-effects IVW methods, we deleted SNPs with *P*-value < 1.00E-5 to reduce heterogeneity. For BD, excluding rs6938026, the heterogeneity (P_Q = 0.010) is reduced. For MDD, the heterogeneity (P_Q = 0.004) is reduced after excluding rs11693221. The heterogeneity of SCZ (P_Q = 0.078) is also reduced after excluding rs6938026 and rs370771.

In addition, we performed a series of pleiotropic tests to further ensure the validation of MR analysis. For MDD, MR-Egger showed that the intercept is not statistically significant (p = 0.131), which indicated that there was no horizontal pleiotropy that existed among IVs. The MR-PRESSO outlier test suggested that rs11693221 (RSS_{obs} = 2.32E-3, p < 0.01) was a potential outlier. We used the MR-PRESSO distortion test and LOO test to test whether the causal effect changed with or without the outlier, but their results were different ($P_{MR\text{-}PRESSO}$ = 0.282 and P_{LOO} = 1.03E-05). Then, even excluding the outlier, the heterogeneity test showed statistically significant heterogeneity (P_Q = 0.004). Therefore, we used the result from random-effects IVW method with the outlier excluded to represent the casual effect of insomnia on MDD (OR = 1.408, 95%CI: 1.209–1.640, p = 1.03E-05) (**Figure 3** and **Supplementary Table S4**).

For BD, MR-Egger showed that the intercept is not statistically significant (p = 0.496). The MR-PRESSO outlier test suggested that rs6938026 (RSS_{obs} = 2.89E-3, p < 0.01) and rs77960 (RSS_{obs} = 2.28E-3, p = 0.013) were the potential outliers. However, the MR-PRESSO distortion test and LOO test indicated that no statistical significance could be identified when excluding the two variants ($P_{MR\text{-}PRESSO}$ = 0.979, P_{LOO1} = 0.147, and P_{LOO1} = 0.622). Though heterogeneity was reduced without the two outliers, we fail to define a statistically significant causal effect for insomnia on BD (**Supplementary Figure S2** and **Supplementary Table S3**).

For SCZ, MR-Egger showed that the intercept is not statistically significant (p = 0.689). The MR-PRESSO outlier test indicated that four SNPs, including rs1456193, rs370771, rs4986172, and rs6938026, were identified as potential outliers. However, the MR-PRESSO distortion test and LOO test indicated that no statistical significance could be identified when excluding the two variants ($P_{MR\text{-}PRESSO}$ = 0.738, P_{LOO1} = 0.172, P_{LOO2} = 0.588, P_{LOO3} = 0.096, and P_{LOO4} = 0.613). Though there was no significant heterogeneity between models with and without outliers, we used the result from the random-effects IVW method with the outliers excluded to represent the casual effect of insomnia on SCZ for caution (OR = 0.752, 95%CI: 0.524–1.079, p = 0.122) (**Supplementary Figure S3** and **Supplementary Table S5**).

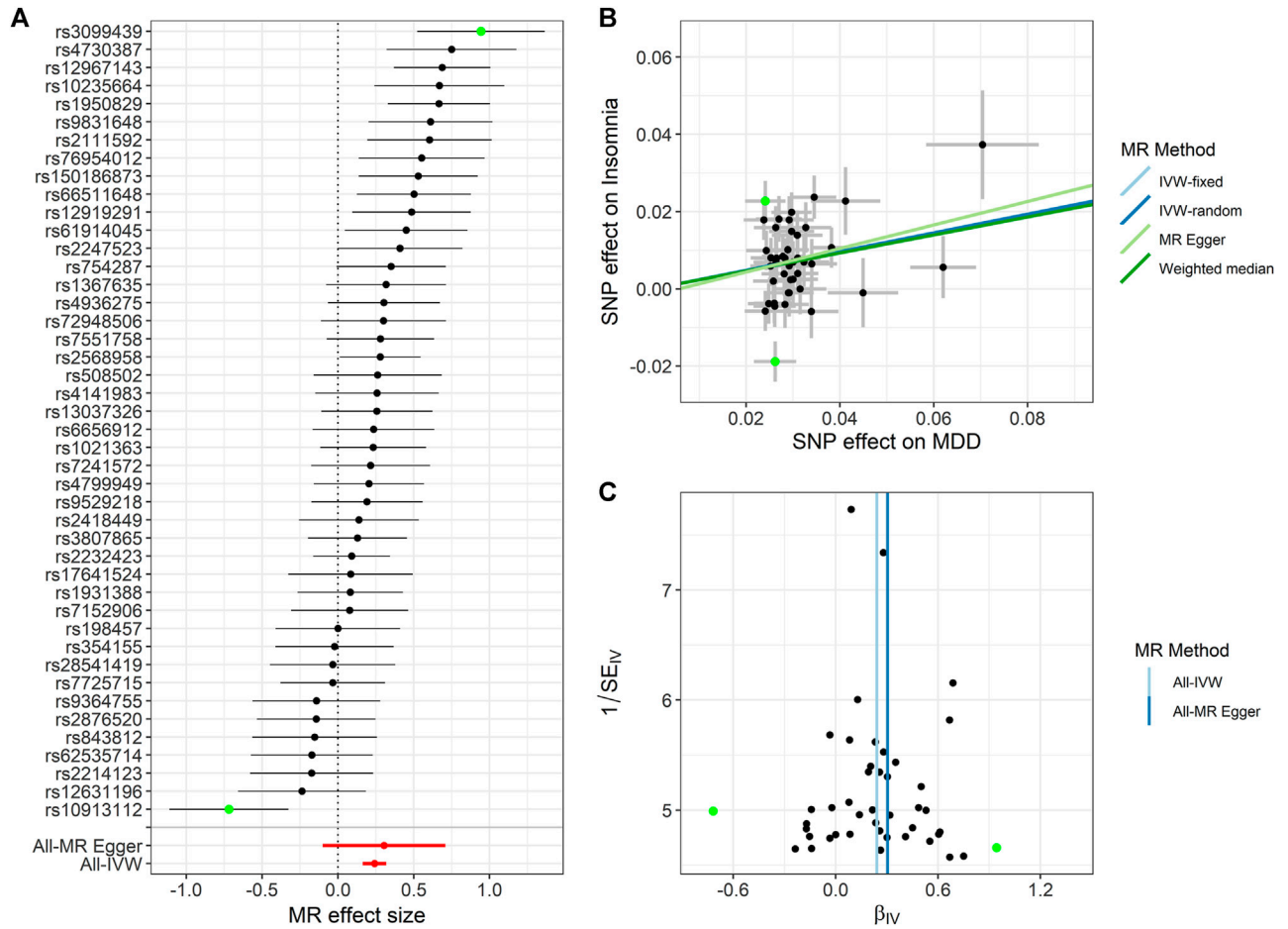

FIGURE 4 | Summary of the reverse-directional Mendelian randomization (MR) analysis for insomnia on major depression (MDD). **(A)** Reverse-directional MR effect size of each instrumental variable (IV), MR-Egger, and inverse variance weighted (IVW). **(B)** Scatter plot of the causal effects of MDD on insomnia. We use vertical and horizontal black lines to show 95% CI of the estimated effect of IVs on MDD (x-axis) and that on insomnia (y-axis), respectively. We use the blue line to show the IVW random-effects model. The potential SNP outliers (rs3099439 and rs10913112) are highlighted in green. **(C)** Funnel plot of the causal effect of MDD on insomnia. Each point represents the estimated causal effect of each IV. The vertical dark blue line represents the causal effect estimate obtained using the MR-Egger method; the light blue line represents the causal effect estimate obtained using the IVW method. The potential outliers (rs3099439 and rs10913112) are highlighted in green.

Reverse-Direction MR Analysis

Following a previous MR analysis (Hartwig et al., 2017; Dong et al., 2021), in order to identify the potential confounding factors that mislead the direction of causal effects, we performed reverse-direction MR (**Figure 4**, **Supplementary Figures S4, S5**). We found that MDD and SCZ have a significantly potential causal association to insomnia, while a potential causal effect for BD on insomnia is not significant. Specifically, using IVW, the estimated OR for MDD and SCZ on insomnia is 1.273 (p = 1.097E-9) and 1.028 (p = 0.004), respectively (**Figure 4** and **Supplementary Figure S4**). The results indicate that the risk of BD and SCZ could increase the risk of insomnia.

DISCUSSION

Using the summary statistics of four traits and reference LD panel from public sources, we performed a two-sample MR analysis to show the causal effects of insomnia on three psychiatric disorders. We found that the causal OR of insomnia on MDD is 1.288, that the reverse direction causal OR of MDD on insomnia is 1.230, and that no statistical significance is defined for insomnia on BD and SCZ. These results were based on several MR methods to guard against potential model misspecifications and is consistent in the estimates of causal effects, suggesting that the findings are convincing.

As we have known, many observational studies aim to explore the associations between insomnia and BD, MDD, and SCZ. A case–control study found that insomnia significantly affected patients with BP with depressive symptoms (OR = 4.17, p = 0.043), and sleep disturbances also predicted manic symptoms (OR = 8.69, p = 0.001) (Palagini et al., 2020). Integrating 21 observational studies for insomnia on DP, Baglioni *et al.* show that the overall OR of insomnia is 2.60 (Baglioni et al., 2011). A cross-sectional study found that the effect size of insomnia-caused symptoms of depression or anxiety is 3.01 (Batalla-Martín et al., 2020).

The abovementioned studies have shown that insomnia is a risk factor to psychiatric disorders. Differently to previous studies, the effect size from MR is directional.

The causal relationship between insomnia and BD, MDD, and SCZ identified in the European population was estimated using the IVs of insomnia in three different outcomes. However, we also recognize that there is still a large amount of unexplainable diversity in the etiology of BD, MDD, and SCZ in European populations. Further research is needed to understand the genetic and environmental factors behind the differences between BD, MDD, and SCZ. Although many studies have confirmed the potential impact of insomnia symptoms on some psychiatric disorders, as mentioned earlier, there is no clear answer yet, and it is not clear whether insomnia has a significant causal effect on these diseases.

Like other MR analyses, our results are not without any drawbacks. First, MR cannot completely exclude all confounding factors because the relationship between exposure and outcome obtained through the observational data used in MR analysis is not a pure relationship between exposure and outcome (Sekula et al., 2016; Ference et al., 2019). In our research, it may be because the sample size of BD and SCZ is relatively small compared with insomnia, and the effect of exposure on the results is relatively weak. The statistical power of MR analysis for certain exposures is limited, resulting in negative results (Pierce and Burgess, 2013). Second, we defined the bidirectional causal association between insomnia and MDD. This plays an important supplement to support the causal association, such that it is hard to detangle the relationship between them using either a cross-sectional study or a MR analysis (Manber and Chambers, 2009; Fang et al., 2019). Nevertheless, our study also provides help for new developments in psychiatric disorder research and new treatment strategies in the future (Hertenstein et al., 2019).

CONCLUSION

The result of the MR and additional analyses shows that insomnia has a positive causal effect on MDD in the European population and provides new evidence of the causal relationship with insomnia on BD and SCZ in European populations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SY designed the study. XZ, XY, and YW performed the datasets quality control. YZ and XY performed the data analysis. PH, XZ, and XY interpreted the analysis results. PH and YZ wrote the draft manuscript. RY and SY revised the article. All the authors accepted the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 81703321) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

ACKNOWLEDGMENTS

We acknowledge the participants and investigators of GWAS-ALTAS for making the summary data publicly available for us.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.763259/full#supplementary-material

REFERENCES

Ayuso-Mateos, J. L., Vázquez-Barquero, J. L., Dowrick, C., Lehtinen, V., Dalgard, O. S., Casey, P., et al. (2001). Depressive Disorders in Europe: Prevalence Figures from the ODIN Study. *Br. J. Psychiatry* 179, 308–316. doi:10.1192/bjp.179.4.308

Baglioni, C., Battagliese, G., Feige, B., Spiegelhalder, K., Nissen, C., Voderholzer, U., et al. (2011). Insomnia as a Predictor of Depression: a Meta-Analytic Evaluation of Longitudinal Epidemiological Studies. *J. Affective Disord.* 135, 10–19. doi:10.1016/j.jad.2011.01.011

Batalla-Martín, D., Belzunegui-Eraso, A., Miralles Garijo, E., Martínez Martín, E., Romaní Garcia, R., Heras, J. S. M., et al. (2020). Insomnia in Schizophrenia Patients: Prevalence and Quality of Life. *Ijerph* 17, 1350. doi:10.3390/ijerph17041350

Bertisch, S. M., Pollock, B. D., Mittleman, M. A., Buysse, D. J., Bazzano, L. A., Gottlieb, D. J., et al. (2018). Insomnia with Objective Short Sleep Duration and Risk of Incident Cardiovascular Disease and All-Cause Mortality: Sleep Heart Health Study. *Sleep* 41. doi:10.1093/sleep/zsy047

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44, 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016a). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40, 304–314. doi:10.1002/gepi.21965

Bowden, J., Del Greco M., F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016b). Assessing the Suitability of Summary Data for Two-Sample Mendelian Randomization Analyses Using MR-Egger Regression: the Role of the I2 Statistic. *Int. J. Epidemiol.* 45, dyw220–1974. doi:10.1093/ije/dyw220

Chen, H.-C., Su, T.-P., and Chou, P. (2013). A Nine-Year Follow-Up Study of Sleep Patterns and Mortality in Community-Dwelling Older Adults in Taiwan. *Sleep* 36, 1187–1198. doi:10.5665/sleep.2884

Davey Smith, G., and Hemani, G. (2014). Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies. *Hum. Mol. Genet.* 23, R89–R98. doi:10.1093/hmg/ddu328

Di Luca, M., Baker, M., Corradetti, R., Kettenmann, H., Mendlewicz, J., Olesen, J., et al. (2011). Consensus Document on European Brain Research. *Eur. J. Neurosci.* 33, 768–818. doi:10.1111/j.1460-9568.2010.07596.x

Dong, S.-S., Zhang, K., Guo, Y., Ding, J.-M., Rong, Y., Feng, J.-C., et al. (2021). Phenome-wide Investigation of the Causal Associations between Childhood BMI and Adult Trait Outcomes: a Two-Sample Mendelian Randomization Study. *Genome Med.* 13, 48. doi:10.1186/s13073-021-00865-3

Dong, Y., and Yang, F. M. (2019). Insomnia Symptoms Predict Both Future Hypertension and Depression. *Prev. Med.* 123, 41–47. doi:10.1016/j.ypmed.2019.02.001

Fang, H., Tu, S., Sheng, J., and Shao, A. (2019). Depression in Sleep Disturbance: A Review on a Bidirectional Relationship, Mechanisms and Treatment. *J. Cel Mol Med* 23, 2324–2332. doi:10.1111/jcmm.14170

Ference, B. A., Ray, K. K., Catapano, A. L., Ference, T. B., Burgess, S., Neff, D. R., et al. (2019). Mendelian Randomization Study of ACLY and Cardiovascular Disease. *N. Engl. J. Med.* 380, 1033–1042. doi:10.1056/NEJMoa1806747

Franzen, P. L., and Buysse, D. J. (2008). Sleep Disturbances and Depression: Risk Relationships for Subsequent Depression and Therapeutic Implications. *Dialogues Clin. Neurosci.* 10, 473–481. doi:10.31887/DCNS.2008.10.4/plfranzen

Gormley, M., Yarmolinsky, J., Dudding, T., Burrows, K., Martin, R. M., Thomas, S., et al. (2021). Using Genetic Variants to Evaluate the Causal Effect of Cholesterol Lowering on Head and Neck Cancer Risk: A Mendelian Randomization Study. *Plos Genet.* 17, e1009525. doi:10.1371/journal.pgen.1009525

Grandner, M. A., Seixas, A., Shetty, S., and Shenoy, S. (2016). Sleep Duration and Diabetes Risk: Population Trends and Potential Mechanisms. *Curr. Diab Rep.* 16, 106. doi:10.1007/s11892-016-0805-8

Hartwig, F. P., Borges, M. C., Horta, B. L., Bowden, J., and Davey Smith, G. (2017). Inflammatory Biomarkers and Risk of Schizophrenia. *JAMA Psychiatry* 74, 1226–1233. doi:10.1001/jamapsychiatry.2017.3191

Hemani, G., Bowden, J., and Davey Smith, G. (2018). Evaluating the Potential Role of Pleiotropy in Mendelian Randomization Studies. *Hum. Mol. Genet.* 27, R195–r208. doi:10.1093/hmg/ddy163

Hertenstein, E., Feige, B., Gmeiner, T., Kienzler, C., Spiegelhalder, K., Johann, A., et al. (2019). Insomnia as a Predictor of Mental Disorders: A Systematic Review and Meta-Analysis. *Sleep Med. Rev.* 43, 96–105. doi:10.1016/j.smrv.2018.10.006

Hou, L., Bergen, S. E., Akula, N., Song, J., Hultman, C. M., Landén, M., et al. (2016). Genome-wide Association Study of 40,000 Individuals Identifies Two Novel Loci Associated with Bipolar Disorder. *Hum. Mol. Genet.* 25, 3383–3394. doi:10.1093/hmg/ddw181

Howard, D. M., Adams, M. J., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., et al. (2019). Genome-wide Meta-Analysis of Depression Identifies 102 Independent Variants and Highlights the Importance of the Prefrontal Brain Regions. *Nat. Neurosci.* 22, 343–352. doi:10.1038/s41593-018-0326-7

Jansen, P. R., Watanabe, K., Watanabe, K., Stringer, S., Skene, N., Bryois, J., et al. (2019). Genome-wide Analysis of Insomnia in 1,331,010 Individuals Identifies New Risk Loci and Functional Pathways. *Nat. Genet.* 51, 394–403. doi:10.1038/s41588-018-0333-3

Javaheri, S., and Redline, S. (2017). Insomnia and Risk of Cardiovascular Disease. *CHEST* 152, 435–444. doi:10.1016/j.chest.2017.01.026

Manber, R., and Chambers, A. S. (2009). Insomnia and Depression: A Multifaceted Interplay. *Curr. Psychiatry Rep.* 11, 437–442. doi:10.1007/s11920-009-0066-1

Manolio, T. A., Rodriguez, L. L., Brooks, L., Abecasis, G., Ballinger, D., Daly, M., et al. (2007). New Models of Collaboration in Genome-wide Association Studies: the Genetic Association Information Network. *Nat. Genet.* 39, 1045–1051. doi:10.1038/ng2127

Morin, C. M., Drake, C. L., Harvey, A. G., Krystal, A. D., Manber, R., Riemann, D., et al. (2015). Insomnia Disorder. *Nat. Rev. Dis. Primers* 1, 15026. doi:10.1038/nrdp.2015.26

Nagel, M., Jansen, P. R., Jansen, P. R., Stringer, S., Watanabe, K., de Leeuw, C. A., et al. (2018). Meta-analysis of Genome-wide Association Studies for Neuroticism in 449,484 Individuals Identifies Novel Genetic Loci and Pathways. *Nat. Genet.* 50, 920–927. doi:10.1038/s41588-018-0151-7

Noyce, A. J., Kia, D. A., Hemani, G., Nicolas, A., Price, T. R., De Pablo-Fernandez, E., et al. (2017). Estimating the Causal Influence of Body Mass index on Risk of Parkinson Disease: A Mendelian Randomisation Study. *Plos Med.* 14, e1002314. doi:10.1371/journal.pmed.1002314

Ong, J. S., and Macgregor, S. (2019). Implementing MR-PRESSO and GCTA-GSMR for Pleiotropy Assessment in Mendelian Randomization Studies from a Practitioner's Perspective. *Genet. Epidemiol.* 43, 609–616. doi:10.1002/gepi.22207

Palagini, L., Miniati, M., Caruso, D., Massa, L., Novi, M., Pardini, F., et al. (2020). Association between Affective Temperaments and Mood Features in Bipolar Disorder II: The Role of Insomnia and Chronobiological Rhythms Desynchronization. *J. Affective Disord.* 266, 263–272. doi:10.1016/j.jad.2020.01.134

Palagini, L., Miniati, M., Marazziti, D., Sharma, V., and Riemann, D. (2021). Association Among Early Life Stress, Mood Features, Hopelessness and Suicidal Risk in Bipolar Disorder: The Potential Contribution of Insomnia Symptoms. *J. Psychiatr. Res.* 135, 52–59. doi:10.1016/j.jpsychires.2020.12.069

Pardiñas, A. F., Holmans, P., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., et al. (2018). Common Schizophrenia Alleles Are Enriched in Mutation-Intolerant Genes and in Regions under strong Background Selection. *Nat. Genet.* 50, 381–389. doi:10.1038/s41588-018-0059-2

Parthasarathy, S., Vasquez, M. M., Halonen, M., Bootzin, R., Quan, S. F., Martinez, F. D., et al. (2015). Persistent Insomnia Is Associated with Mortality Risk. *Am. J. Med.* 128, 268–275.e262. doi:10.1016/j.amjmed.2014.10.015

Paternoster, L., Tilling, K., and Davey Smith, G. (2017). Genetic Epidemiology and Mendelian Randomization for Informing Disease Therapeutics: Conceptual and Methodological Challenges. *Plos Genet.* 13, e1006944. doi:10.1371/journal.pgen.1006944

Peng, J., Guan, J., Hui, W., and Shang, X. (2021a). A Novel Subnetwork Representation Learning Method for Uncovering Disease-Disease Relationships. *Methods* 192, 77–84. doi:10.1016/j.ymeth.2020.09.002

Peng, J., Wang, Y., Guan, J., Li, J., Han, R., Hao, J., et al. (2021b). An End-To-End Heterogeneous Graph Representation Learning-Based Framework for Drug-Target Interaction Prediction. *Brief. Bioinform.* 22. doi:10.1093/bib/bbaa430

Peng, J., Xue, H., Wei, Z., Tuncali, I., Hao, J., and Shang, X. (2020). Integrating Multi-Network Topology for Gene Function Prediction Using Deep Neural Networks. *Brief. Bioinform.* 22, 2096–2105. doi:10.1093/bib/bbaa036

Pierce, B. L., and Burgess, S. (2013). Efficient Design for Mendelian Randomization Studies: Subsample and 2-sample Instrumental Variable Estimators. *Am. J. Epidemiol.* 178, 1177–1184. doi:10.1093/aje/kwt084

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'donovan, M. C., Sullivan, P. F., et al. (2009). Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. *Nature* 460, 748–752. doi:10.1038/nature08185

Riemann, D., and Voderholzer, U. (2003). Primary Insomnia: a Risk Factor to Develop Depression?. *J. Affective Disord.* 76, 255–259. doi:10.1016/s0165-0327(02)00072-1

Ripke, S., O'dushlaine, C., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., et al. (2013). Genome-wide Association Analysis Identifies 13 New Risk Loci for Schizophrenia. *Nat. Genet.* 45, 1150–1159. doi:10.1038/ng.2742

Roth, T. (2007). Insomnia: Definition, Prevalence, Etiology, and Consequences. *J. Clin. Sleep Med.* 3, S7–S10. doi:10.5664/jcsm.26929

Ruderfer, D. M., Ripke, S., Mcquillin, A., Boocock, J., Stahl, E. A., Pavlides, J. M. W., et al. (2018). Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* 173, 1705–e16.e1716. doi:10.1016/j.cell.2018.05.046

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., De Leeuw, C. A., et al. (2018). Genome-wide Association Meta-Analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence. *Nat. Genet.* 50, 912–919. doi:10.1038/s41588-018-0152-6

Sekula, P., Del Greco M, F., Pattaro, C., and Köttgen, A. (2016). Mendelian Randomization as an Approach to Assess Causality Using Observational Data. *Jasn* 27, 3253–3265. doi:10.1681/asn.2016010098

Sklar, P., Ripke, S., Scott, L. J., Andreassen, O. A., Cichon, S., Craddock, N., et al. (2011). Large-scale Genome-wide Association Analysis of Bipolar Disorder Identifies a New Susceptibility Locus Near ODZ4. *Nat. Genet.* 43, 977–983. doi:10.1038/ng.943

Sleiman, P. M., and Grant, S. F. (2010). Mendelian Randomization in the Era of Genomewide Association Studies. *Clin. Chem.* 56, 723–728. doi:10.1373/clinchem.2009.141564

Smith, E. N., Bloss, C. S., Badner, J. A., Barrett, T., Belmonte, P. L., Berrettini, W., et al. (2009). Genome-wide Association Study of Bipolar Disorder in European American and African American Individuals. *Mol. Psychiatry* 14, 755–763. doi:10.1038/mp.2009.43

Stummer, L., Markovic, M., and Maroney, M. (2018). Pharmacologic Treatment Options for Insomnia in Patients with Schizophrenia. *Medicines* 5, 88. doi:10.3390/medicines5030088

Tian, D., Wang, P., Tang, B., Teng, X., Li, C., Liu, X., et al. (2020). GWAS Atlas: a Curated Resource of Genome-wide Variant-Trait Associations in Plants and Animals. *Nucleic Acids Res.* 48, D927–d932. doi:10.1093/nar/gkz828

Tylee, D. S., Sun, J., Hess, J. L., Tahir, M. A., Sharma, E., Malik, R., et al. (2018). Genetic Correlations Among Psychiatric and Immune-Related Phenotypes Based on Genome-wide Association Data. *Am. J. Med. Genet.* 177, 641–657. doi:10.1002/ajmg.b.32652

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50, 693–698. doi:10.1038/s41588-018-0099-7

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans. *Bioinformatics* 26, 2190–2191. doi:10.1093/bioinformatics/btq340

Winkelman, J. W. (2015). Insomnia Disorder. *N. Engl. J. Med.* 373, 1437–1444. doi:10.1056/NEJMcp1412740

Zeng, P., Wang, T., Zheng, J., and Zhou, X. (2019). Causal Association of Type 2 Diabetes with Amyotrophic Lateral Sclerosis: New Evidence from Mendelian Randomization Using GWAS Summary Statistics. *BMC Med.* 17, 225. doi:10.1186/s12916-019-1448-9

Zeng, P., and Zhou, X. (2019). Causal Effects of Blood Lipids on Amyotrophic Lateral Sclerosis: a Mendelian Randomization Study. *Hum. Mol. Genet.* 28, 688–697. doi:10.1093/hmg/ddy384

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer PZ declared a past co-authorship with one of the authors SY to the handling editor.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 October 2021
doi: 10.3389/fgene.2021.761629

Developing an Embedding, Koopman and Autoencoder Technologies-Based Multi-Omics Time Series Predictive Model (EKATP) for Systems Biology research

*Suran Liu[1†], Yujie You[1†], Zhaoqi Tong[2] and Le Zhang[1]**

[1]College of Computer Science, Sichuan University, Chengdu, China, [2]College of Software Engineering, Sichuan University, Chengdu, China

It is very important for systems biologists to predict the state of the multi-omics time series for disease occurrence and health detection. However, it is difficult to make the prediction due to the high-dimensional, nonlinear and noisy characteristics of the multi-omics time series data. For this reason, this study innovatively proposes an Embedding, Koopman and Autoencoder technologies-based multi-omics time series predictive model (EKATP) to predict the future state of a high-dimensional nonlinear multi-omics time series. We evaluate this EKATP by using a genomics time series with chaotic behavior, a proteomics time series with oscillating behavior and a metabolomics time series with flow behavior. The computational experiments demonstrate that our proposed EKATP can substantially improve the accuracy, robustness and generalizability to predict the future state of a time series for multi-omics data.

Keywords: multi-omics, time series, embedding, Koopman, deep learning

OPEN ACCESS

Edited by:
Jiajie Peng, Northwestern Polytechnical University, China

Reviewed by:
Renchu Guan, Jilin University, China
Xueming Liu, Huazhong University of Science and Technology, China

****Correspondence:***
Le Zhang
zhangle06@scu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *20 August 2021*
Accepted: *27 September 2021*
Published: *26 October 2021*

Citation:
Liu S, You Y, Tong Z and Zhang L (2021) Developing an Embedding, Koopman and Autoencoder Technologies-Based Multi-Omics Time Series Predictive Model (EKATP) for Systems Biology research. Front. Genet. 12:761629. doi: 10.3389/fgene.2021.761629

INTRODUCTION

Currently, the prediction of multi-omics time series states is one of the trending areas in systems biology research (Zhang et al., 2019a). In particular, the development of high-throughput technology (Soon et al., 2013) has produced a large-scale time series multi-omics state (Liang and Kelemen 2017a), including genomics (Lockhart and Winzeler 2000), proteomics (Tyers and Mann, 2003), metabolomics (Weckwerth 2003) and more. Previous studies usually employed differential equation (Eisenhammer et al., 1991; Zhang et al., 2016; Zhang and Zhang 2017; Liu G.-D. et al., 2020) based models to abstract and formalise multi-omics time series data (Bianconi et al., 2020). Then, it became possible to explore the time-varying connections and predict their future state (Ji et al., 2017) by solving these differential equations. In particular, predicting multi-omics time series states can not only discover dynamic information for biological entities, such as genes, proteins and metabolites, but also explore complicated biological interactions and the pathogenesis of diseases (Liang and Kelemen, 2017b).

However, a multi-omics time series usually has high dimensions (Perez-Riverol et al., 2017), complicated interaction relationships (Fischer 2008) and inevitable noise (Fischer 2008; Tsimring 2014). Thus, when we employ differential equations to model the multi-omics time series state, it is hard for us to solve these equations due to their high dimensionality and nonlinear characteristics (Bianconi et al., 2020). For these reasons, the way to predict the future state of a multi-omics time

series by solving these complicated nonlinear differential equations has already become challenging work.

Recently, future state prediction for a multi-omics time series has been widely studied by computational biologists. For genomic studies, we usually use a gene expression time series to develop gene regulatory networks (Davidson and Levin 2005; Zhang et al., 2018; Xiao et al., 2020; Zhang et al., 2020; Xiao et al., 2021; Zhang et al., 2021a). However, since the gene regulatory network is a complex high-dimensional nonlinear system (Zhang et al., 2012a), it often produces chaotic phenomena (Levnajić and Tadić 2010), which not only play an important role in maintaining stable gene expression patterns (Sevim and Rikvold 2008) but also are closely related to the occurrence of diseases (Suzuki et al., 2016). Usually, we employ the Lorentz system (Lorenz 1963) to describe the chaotic phenomenon. However, it is inaccurate to predict the future state of genomics time series with nonlinear complicated interactions because the Lorentz system is not good at processing nonlinear complicated interactions (Lai et al., 2018). Currently, delay embedding theory (Sauer et al., 1991; Holmes et al., 2012) is commonly used to transform the spatial information (complicated interactions) into temporal information (the future state of the time series (Chen et al., 2020)) for dimensional reduction (Gao et al., 2017; Li et al., 2017; Xia et al., 2017; Zhang et al., 2019b; Zhang et al., 2019c; Wu et al., 2020; You et al., 2020; Zhang et al., 2021b), whereas Koopman theory (Koopman, 1931) can switch the nonlinear system into a linear system to reduce computing cost. Therefore, our first research question asks if we can develop such a time series predictive model that integrates the Lorentz system with delay embedding and Koopman theory to accurately predict the future state of genomics time series with chaotic behavior.

For proteomics studies, we usually use proteomic time series data to infer protein–protein interactions (PPIs) (Wu et al., 2009). Currently, we employ mass spectrometry technology (Mann et al., 2001) to obtain proteomics time series data. However, since it is unstable to have time-course experimental data by mass spectrometry technology, proteomics time series data are prone to oscillating behavior (Iuchi et al., 2018). Previously, we employed a nonlinear pendulum system (Hirsch 1974) to describe the oscillation behavior, though it was subjected to overfitting under a strong noise environment. Since the conjugate form of delay embedding (Sauer et al., 1991; Holmes et al., 2012) can ensure the reversibility of the time series predictive model (Chen et al., 2020) and reduce the impact of noise on prediction to a certain extent, our second research question asks if we can develop such a time series predictive model that can integrate a nonlinear pendulum system with delay embedding to accurately predict the state of proteomics time series with oscillating behavior.

For metabolomics studies, we usually use metabolic time series data that represent the flow behavior of biological fluids (serum, cerebrospinal fluid, etc.) to discover key metabolites in biological fluids (Zhang et al., 2012b). A previous study (Noack et al., 2003) always employed a nonlinear biological fluid system to describe metabolic time series data. However, because most nonlinear fluid flow systems have high dimensions (Lusch et al., 2018), we not only have difficulty selecting features from high-dimensional metabolic time series data but also impede progress because of time-consuming computing (Wang et al., 2021). Currently, since neural networks (Wang et al., 2014) can decrease the computing cost (Song et al., 2017) by dimensional reduction for time series data (Hinton and Salakhutdinov, 2006), our third research question asks if we can develop such a time series predictive model that integrates a nonlinear fluid flow system with a neural network to predict the future state of the metabolomics time series accurately and quickly with flow behaviour.

To answer the above three research questions, this study innovatively develops an Embedding, Koopman and Autoencoder technologies-based multi-omics time series predictive model (EKATP) to predict the future state of the time series for the corresponding genomics, proteomics and metabolomics datasets. Compared with previous approaches (Lusch et al., 2018; Azencot et al., 2020), the contributions of the study are summarised as follows. First, we select key features from a high-dimensional nonlinear state by integrating a neural network with the delay embedding theory. Second, we switch the nonlinear system with a linear system to reduce the computing cost by the Koopman theory. Finally, we develop a neural network and delay embedding theory-based model for reversible mapping between a high-dimensional nonlinear system and a low-dimensional linear system, thereby improving the accuracy and robustness of prediction.

The rest of the manuscript is organised as follows. *Related Works* mainly describes the related work for Autoencoder, delay embedding theory and Koopman theory. *Materials and Methods* introduces the architecture of the EKATP and the related procedure. *Experiments* describes the computational experiments. Finally, we conclude the study and discuss the future work.

RELATED WORKS

Supplementary Presentation S1 details the related theory and existing research of the Autoencoder, delay embedding theory and Koopman theory.

MATERIALS AND METHODS

Figure 1 describes the workflow of the EKATP.

Problem and Definitions

Given a set of high-dimensional nonlinear multi-omics time series states $\boldsymbol{F} = (\boldsymbol{F^1}, \boldsymbol{F^2}, \ldots, \boldsymbol{F^T})$, where T represents the total step, the time series state at t can be described as $\boldsymbol{F^t} = (\boldsymbol{f_1^t}, \boldsymbol{f_2^t}, \ldots \boldsymbol{f_n^t})'$, where n represents the dimension of the time series state, " ", as the transpose of a vector. Our goal is to predict the future state of the multi-omics time series. Next, we detail how to develop an EKATP as follows.

Autoencoding Observations

Since an EKATP is based on the Autoencoder framework, we employ **Eq. 1** to define the objective function for Autoencoder (L_{id}).

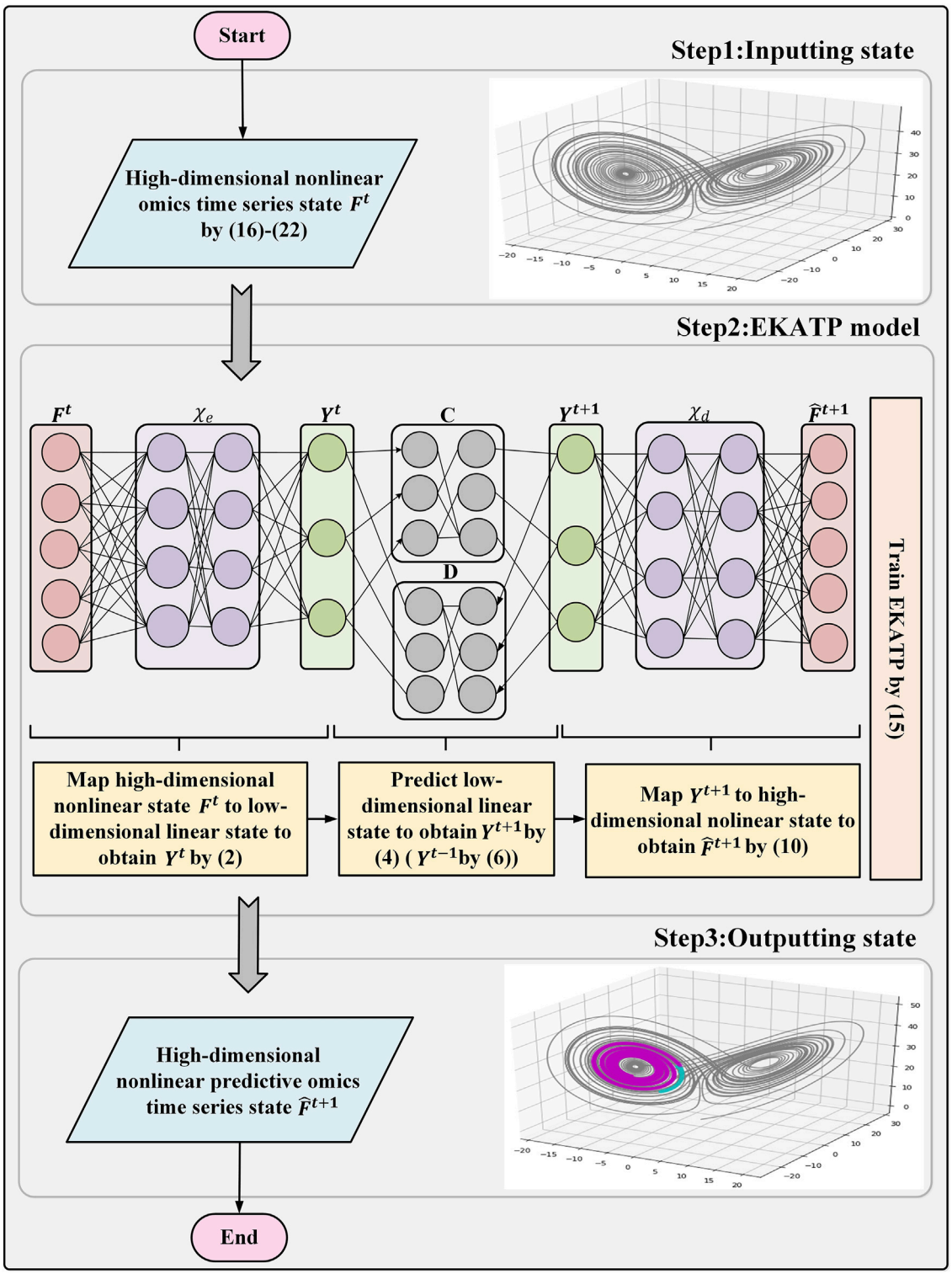

FIGURE 1 | EKATP workflow.

$$L_{id} = \left\| \hat{\boldsymbol{F}}^{t} - \boldsymbol{F}^{t} \right\|_{\mathrm{MSE}} \tag{1}$$

Here, $\hat{\boldsymbol{F}}^{t}$ is the reconstructed high-dimensional time series state according to encoder (χ_e) and decoder (χ_d) of Autoencoder (**Supplementary Presentation S1**). $\|\cdot\|_{\mathrm{MSE}}$ denotes the mean squared error (MSE), which presents the expected value of the square of the difference between the predicted value and the true value. This loss function term enables us to construct an Autoencoder model that satisfies $\chi_d \circ \chi_e \approx id$, the identity.

Delay Embedding

According to the description in the delay embedding theory (**Supplementary Presentation S1**), we employ χ_e of the Autoencoder to approximate the delay embedding Φ, mapping the high-dimensional nonlinear input time series state $\boldsymbol{F}^{t}$ back to the low-dimensional time series state $\boldsymbol{Y}^{t}$ by **Eq. 2**,

$$\boldsymbol{Y}^{t} = \left(y^{t}, y^{t+1}, ..., y^{t+L-1}\right)' = \chi_e\left(\boldsymbol{F}^{t}\right). \tag{2}$$

where L represents the dimension of the low-dimensional time series state. Similarly, the inverse mapping χ_d of mapping χ_e is used to approximate the conjugate form of delay embedding Φ, mapping the low-dimensional time series state back to the high-dimensional time series state by **Eq. 3**.

$$\hat{\boldsymbol{F}}^{t} = \left(\hat{f}_1^{t}, \hat{f}_2^{t}, \ldots \hat{f}_n^{t}\right)' = \chi_d\left(\boldsymbol{Y}^{t}\right). \tag{3}$$

Linearized Representation of the Koopman Operator

Based on the Koopman theory discussed by **Supplementary Presentation S1**, we construct the finite dimensional linear matrix C (and matrix D) to compute the forward (and backward) low-dimensional time series state. **Equation 4** shows how to realize the forward prediction for low-dimensional time series state $\boldsymbol{Y}^{t}$ to obtain $\boldsymbol{Y}^{t+1}$.

$$\boldsymbol{Y}^{t+1} = C\boldsymbol{Y}^{t}. \tag{4}$$

Equation 4 can be expanded by **Eq. 5**.

$$\begin{bmatrix} y^{t+1} \\ y^{t+2} \\ y^{t+3} \\ \ldots \\ y^{t+L-1} \\ y^{t+L} \end{bmatrix} = \begin{bmatrix} 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & 1 \\ a_1 & a_2 & \ldots & a_{L-1} & a_L \end{bmatrix} \begin{bmatrix} y^{t} \\ y^{t+1} \\ y^{t+2} \\ \ldots \\ y^{t+L-2} \\ y^{t+L-1} \end{bmatrix}. \tag{5}$$

Here, a_i is the estimated parameter that needs training, and $a_1 \neq 0$. **Equation 6** shows how to realize the backward prediction for a low-dimensional time series state $\boldsymbol{Y}^{t}$ to obtain $\boldsymbol{Y}^{t-1}$.

$$\boldsymbol{Y}^{t-1} = D\boldsymbol{Y}^{t}. \tag{6}$$

Equation 6 can be expanded by **Eq. 7**.

$$\begin{bmatrix} y^{t-1} \\ y^{t} \\ y^{t+1} \\ \ldots \\ y^{t+L-3} \\ y^{t+L-2} \end{bmatrix} = \begin{bmatrix} b_1 & b_2 & \ldots & b_{L-1} & b_L \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \begin{bmatrix} y^{t} \\ y^{t+1} \\ y^{t+2} \\ \ldots \\ y^{t+L-2} \\ y^{t+L-1} \end{bmatrix}. \tag{7}$$

Here, b_i is the estimated parameter that needs training, and $b_L \neq 0$. Our goal is to optimise the parameters of the linear matrix C (and D) of **Eqs 5**, **7** by model training.

Forward and Backward Prediction

We make the k-steps forward prediction by **Eq. 8** and backward prediction by **Eq. 9** for the state of the low-dimensional time series $\boldsymbol{Y}^{t}$. After that, χ_d is used to map the low-dimensional predictive time series state back to the high-dimensional predictive time series state by **Eq. 10**,

$$\boldsymbol{Y}^{t+k} = C^{k}\boldsymbol{Y}^{t}. \tag{8}$$

$$\boldsymbol{Y}^{t-k} = D^{k}\boldsymbol{Y}^{t}. \tag{9}$$

$$\hat{\boldsymbol{F}}^{t\pm k} = \chi_d\left(\boldsymbol{Y}^{t\pm k}\right). \tag{10}$$

where $\boldsymbol{Y}^{t+k}$ and $\boldsymbol{Y}^{t-k}$ represent the low-dimensional state after k steps of forward and backward prediction, respectively. $\hat{\boldsymbol{F}}^{t\pm k}$ represents the predictive high-dimensional nonlinear state.

Equations 11, **12** define the loss function of forward prediction (L_{fwd}) and backward prediction (L_{bwd}) to minimize the difference between the high-dimensional predictive value and true states at each step, respectively.

$$L_{\mathrm{fwd}} = \frac{1}{k}\sum_{s=1}^{k}\left\|\hat{\boldsymbol{F}}^{t+s} - \boldsymbol{F}^{t+s}\right\|_{\mathrm{MSE}}. \tag{11}$$

$$L_{\mathrm{bwd}} = \frac{1}{k}\sum_{s=1}^{k}\left\|\hat{\boldsymbol{F}}^{t-s} - \boldsymbol{F}^{t-s}\right\|_{\mathrm{MSE}}. \tag{12}$$

Equation 13 defines the loss function (L_{idy}) to minimize the difference between the predictive low-dimensional state obtained by the C and D matrices and defines such a low-dimensional state that is mapped from the true high-dimensional state by mapping χ_e.

$$L_{\mathrm{idy}} = \frac{1}{k}\sum_{s=1}^{k}\left[\left\|C^{s}\chi_e\left(\boldsymbol{F}^{t}\right) - \chi_e\left(\boldsymbol{F}^{t+s}\right)\right\|_{\mathrm{MSE}} + \left\|D^{s}\chi_e\left(\boldsymbol{F}^{t}\right) - \chi_e\left(\boldsymbol{F}^{t-s}\right)\right\|_{\mathrm{MSE}}\right]. \tag{13}$$

Additionally, we employ loss function (L_{con}) by **Eq. 14** to train the parameters a_i and b_i in the matrices C and D, respectively.

$$L_{\mathrm{con}} = \frac{1}{k}\sum_{s=1}^{k}\left[\left\|\chi_d\left(D^{s}C^{s}\boldsymbol{Y}^{t}\right) - \boldsymbol{F}^{t}\right\|_{\mathrm{MSE}} + \left\|\chi_d\left(C^{s}D^{s}\boldsymbol{Y}^{t}\right) - \boldsymbol{F}^{t}\right\|_{\mathrm{MSE}}\right] \tag{14}$$

Parameter Estimation for the EKATP

Equation 15 optimizes the key parameters for the EKATP by minimizing L.

$$L = \lambda_{id}L_{id} + \lambda_{fwd}L_{fwd} + \lambda_{bwd}L_{bwd} + \lambda_{idy}L_{idy} + \lambda_{con}L_{con}. \quad (15)$$

Here, λ_{id}, λ_{fwd}, λ_{bwd}, λ_{idy} and λ_{con} are user-defined hyperparameters.

EXPERIMENTS

This section evaluates the predictability of the proposed EKATP for high-dimensional nonlinear multi-omics datasets by comparing it with recurrent neural networks (RNNs) (Jiang and Lai, 2019), long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), dynamic Autoencoder (DAE) (Lusch et al., 2018) and Koopman Autoencoder (KAE) (Azencot et al., 2020). The detailed experimental setup is listed in **Supplementary Presentation S2**. In addition, we detail the workflow chart and list the related pseudocode in **Supplementary Figure S1**; **Supplementary Presentation S3**.

Genomics

We usually employ the chaotic Lorentz system (Lorenz, 1963) to describe a gene expression time series with a low-dimensional manifold (Sauer et al., 1991) by **Eq. 16**,

$$\begin{cases} x_{t+1} = x_t + h(\eta(y_t - z_t)) \\ y_{t+1} = y_t + h(x_t(\rho - z_t) - y_t), \\ z_{t+1} = z_t + h(x_t y_t - \beta z_t) \end{cases} \quad (16)$$

where η and ρ represent the Prandtl and Rayleigh numbers, respectively. β is related to geometry, and t represents time. h represents the level of the complicated nonlinear system. When h is greater, the nonlinear relationship between genes becomes more complicated.

Since gene expression time series contains considerable noise, we employ white Gaussian noise (Li et al., 2017) to simulate the noise by **Eq. 17**,

$$\begin{cases} \tilde{x} = x + \varepsilon_x \\ \tilde{y} = y + \varepsilon_y \\ \tilde{z} = z + \varepsilon_z \end{cases}. \quad (17)$$

where $\tilde{x}$, $\tilde{y}$ and $\tilde{z}$ represent data with noise. ε_x, ε_y and ε_z represent the white Gaussian noise for x, y and z by normal distributions $N(0, \sigma^2)$ with a zero mean and a standard deviation σ. The standard deviation σ is referred to as the noise intensity.

Here, we describe how to obtain a high-dimensional gene expression time series with a low-dimensional manifold as follows. First, we generate the three-dimensional time series $\boldsymbol{V} = (\boldsymbol{V}^1, \boldsymbol{V}^2, \ldots, \boldsymbol{V}^T) \in \mathbb{R}^3$ (T is the total step), which is listed in **Supplementary Tables S1.1, S1.2, S1.3**. Next, we develop a random orthogonal transformation (Anderson et al., 1987) matrix $\boldsymbol{P} \in \mathbb{R}^{96\times3}$. Finally, we map the state of a 3-dimensional time series onto the state of a 96-dimensional time series by **Eq. 18** to simulate a high-dimensional gene expression time series $\boldsymbol{F} = (\boldsymbol{F}^1, \boldsymbol{F}^2, \ldots, \boldsymbol{F}^T) \in \mathbb{R}^{96}$ with a 3-dimensional manifold, which is listed in **Supplementary Tables S1.4, S1.5, S1.6**.

$$\boldsymbol{F} = \boldsymbol{P}\boldsymbol{V}. \quad (18)$$

To prove the accuracy and robustness of the EKATP, we generate a small-scale system containing T = 1,050 steps and choose the last 50 steps to visualize the predictive power of the EKATP.

Figure 2 shows the predictive error in the range of 50 steps under different initial conditions and environments. Detailed information is listed in **Supplementary Tables S1.7, S1.8**; **Supplementary Presentation S4**.

Figures 2A,C demonstrates that the EKATP not only has less of a predictive error than the existing methods under a clean environment (σ=0.00) but also has a stable predictive error when the complexity h increases from 0.003 to 0.006. In particular, with the increase in predictive steps, the predictive error of the EKATP increases slower than that of the existing methods.

Figures 2B,D shows that the EKATP not only has less of a predictive error than previous methods under a noisy environment (σ=0.01) but also has a predictive error that slightly fluctuates when h increases from 0.003 to 0.006. Moreover, after 25 steps, the predictive error of the EKATP increases much slower than that of the existing methods.

Figure 2 indicates that the EKATP has greater predictive accuracy and robustness than excitation methods in clean and noisy environments.

To further prove the generalizability of the EKATP, we generate a large-scale system containing T = 15,000 steps under the condition of h = 0.003 and σ = 0.00. After that, we randomly choose three different time periods to train and test the model as follows, the procedure of which is detailed in **Supplementary Table S1.9**.

First, since the 3-dimensional time series state and 96-dimensional time series state are diffeomorphic (Sauer et al., 1991), which is indicated by the data preprocessing procedure, it implies that the mapping between these two time series is reversible. Here, we map the 96-dimensional gene expression predictive results onto a 3-dimensional space by orthogonal inverse transformation (Anderson et al., 1987) to visualize the predictive result of the EKATP.

Figures 3A,B,C demonstrates that the predictive results of the EKATP are close to the true value for different periods of a time series. **Figure 3** shows that the EKATP can accurately predict the gene expression time series at different periods, implying that it has a strong generalizability, even in a very complicated nonlinear environment.

Proteomics

We always use a nonlinear pendulum model (Hirsch, 1974) with oscillatory behaviour to describe a proteomics time series with a low-dimensional manifold (Sauer et al., 1991) by **Eq. 19**,

$$\begin{cases} \dfrac{d^2\theta}{dt^2} + \dfrac{g}{l}\sin\theta = 0 \\ \theta(t_0) = h \end{cases}. \quad (19)$$

where l, g and t denote the length, gravity and time, respectively. h denotes the initial value of θ, which represents the level of the complicated nonlinear system. When h is greater, the nonlinear relationship between proteins becomes more complicated.

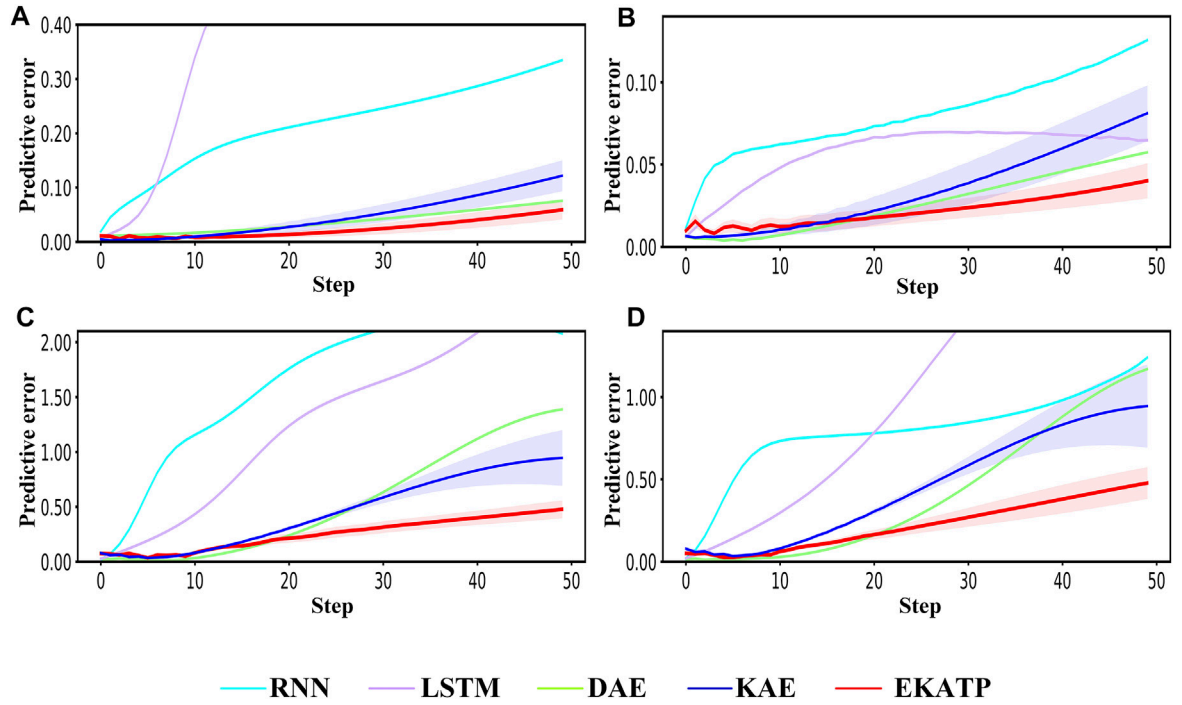

FIGURE 2 | Comparison among the RNN, LSTM, DAE, KAE and EKATP. The abscissa represents the step, and the ordinate represents the predictive error. **(A)** The initial conditions are $h = 0.003$ and $\sigma = 0.00$. **(B)** The initial conditions are $h = 0.003$ and $\sigma = 0.01$. **(C)** The initial conditions are $h = 0.006$ and $\sigma = 0.00$. **(D)** The initial conditions are $h = 0.006$ and $\sigma = 0.01$.

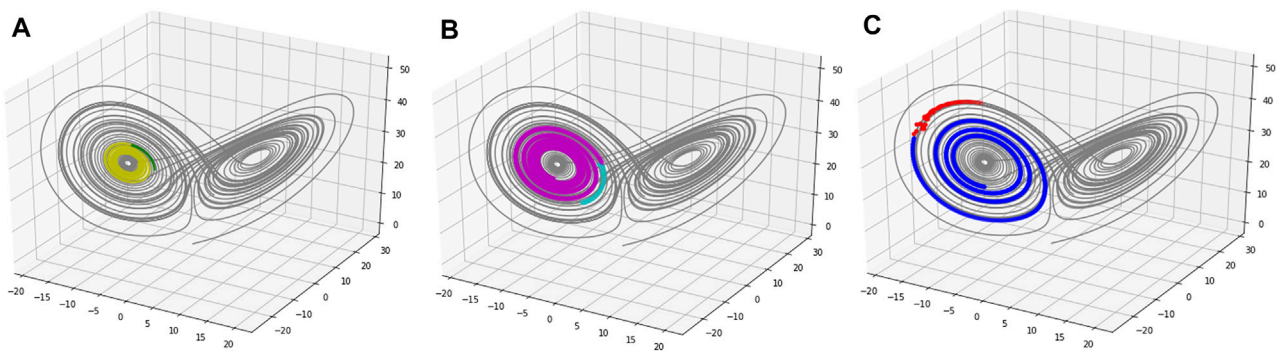

FIGURE 3 | The 50-step predictive trajectories of the EKATP are under initial conditions $h = 0.003$ and $\sigma = 0.00$. Grey colors represent full true data. **(A)** This is the predictive situation of the first period. Yellow and green colors represent true and predictive data, respectively. **(B)** This is the predictive situation of the second period. Purple and cyan colors represent true and predictive data, respectively. **(C)** This is the predictive situation of the third period. Blue and red colors represent true and predictive data, respectively.

Since a considerable amount of noise exists in a protein time series, we employ white Gaussian noise (Li et al., 2017) to describe it by **Eq. 20**,

$$\begin{cases} \tilde{\theta} = \theta + \varepsilon_{\theta} \\ \tilde{\dot{\theta}} = \dot{\theta} + \varepsilon_{\dot{\theta}} \end{cases}, \tag{20}$$

where $\tilde{\theta}$ and $\tilde{\dot{\theta}}$ represent data with noise. ε_{θ} and $\varepsilon_{\dot{\theta}}$ represent the noise Gaussian terms for θ and $\dot{\theta}$ by normal distributions $N(0, \sigma^2)$ with a zero mean and a standard deviation σ.

Here, we describe how to obtain a high-dimensional proteomics time series with a low-dimensional manifold. First, we generate the 2-dimensional time series $\boldsymbol{V} = (\boldsymbol{V}^1, \boldsymbol{V}^2, \ldots, \boldsymbol{V}^T) \in \mathbb{R}^2$, which is listed in **Supplementary Tables S2.1, S2.2**. Next, we develop a random orthogonal transformation (Anderson et al., 1987) matrix $\boldsymbol{P} \in \mathbb{R}^{64\times 2}$. Finally, we map the state of a 2-dimensional time series onto the state of a 64-dimensional time series by **Eq. 18** to simulate a

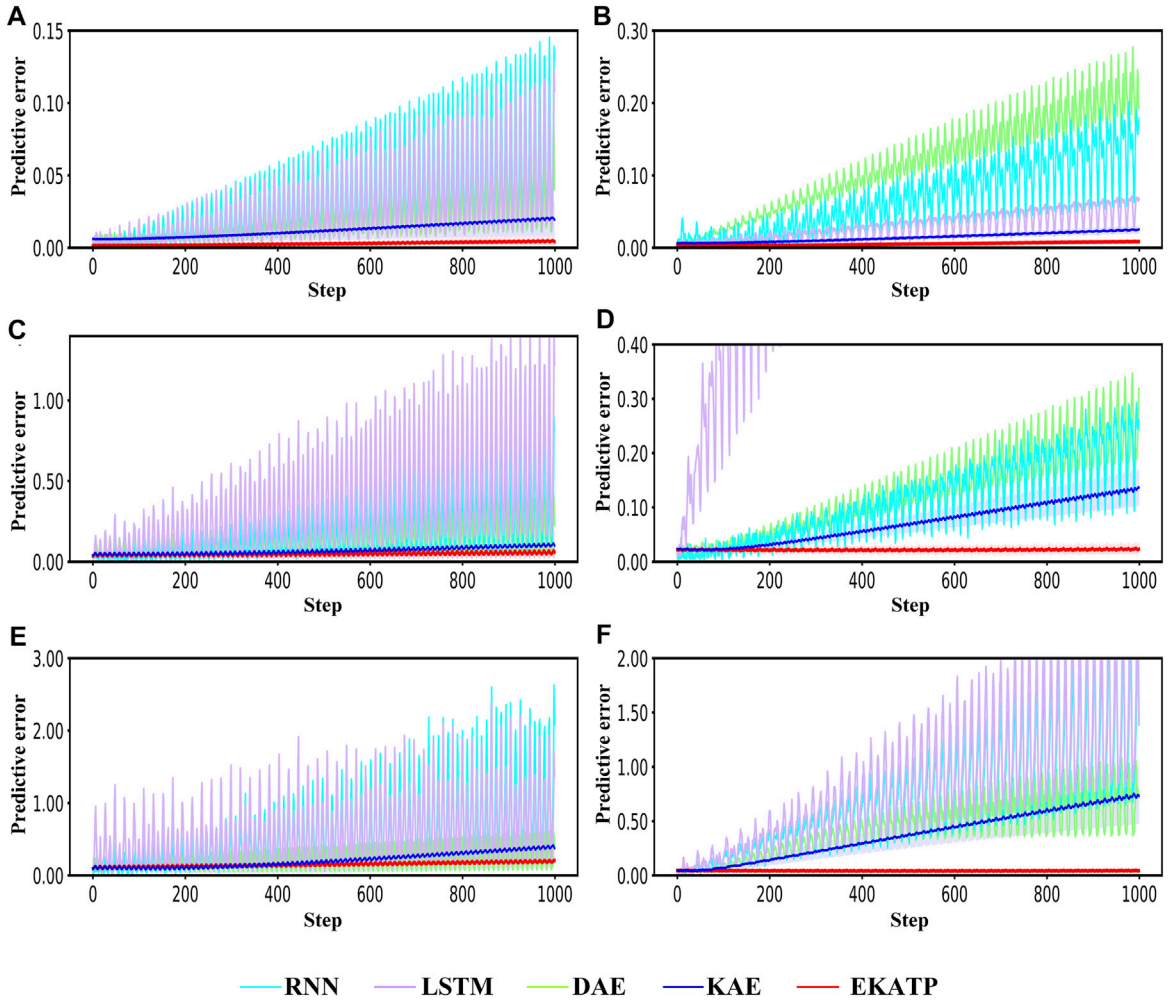

FIGURE 4 | Comparison with the RNN, LSTM, DAE, KAE and EKATP. The abscissa represents the step, and the ordinate represents the predictive error. **(A)** The initial conditions are $h = 0.8$ and $\sigma = 0.00$. **(B)** The initial conditions are $h = 2.4$ and $\sigma = 0.00$. **(C)** The initial conditions are $h = 0.8$ and $\sigma = 0.03$. **(D)** The initial conditions are $h = 2.4$ and $\sigma = 0.03$. **(E)** The initial conditions are $h = 0.8$ and $\sigma = 0.08$. **(F)** The initial conditions are $h = 2.4$ and $\sigma = 0.08$.

high-dimensional proteomics time series $\boldsymbol{F} = (\boldsymbol{F}^1, \boldsymbol{F}^2, \ldots, \boldsymbol{F}^T) \in \mathbb{R}^{64}$ with a 2-dimensional manifold, which is listed in **Supplementary Tables S2.3, S2.4**.

To prove the accuracy and robustness of the EKATP, we generate a system containing $T = 1{,}600$ steps and choose the last 1,000 steps to visualize the predictability for the EKATP.

Figure 4 shows that the EKATP can effectively predict a proteomic time series under clean and noisy environments within 1,000 steps, the details of which are listed in **Supplementary Tables S2.5, S2.6**; **Supplementary Presentation S4**.

Figures 4A,B shows that the EKATP not only has less of a predictive error under a clean environment (σ=0.00) than the existing methods but also maintains a smaller predictive error when h increases from 0.8 to 2.4. Moreover, the predictive error of the EKATP increases much slower than that of the existing methods when the predictive step increases.

Figures 4C,D demonstrates that the EKATP has less of a predictive error under a noise environment (σ=0.03) than the existing methods. When h increases from 0.8 to 2.4, the predictive error of the EKATP remains stable. In particular, with the increase in predictive steps, the predictive error of the EKATP increases much slower than that of the existing methods.

Figures 4E,F indicates that the EKATP not only has less of a predictive error under a noise environment (σ=0.08) than the existing methods but also has a predictive error of the EKATP that remains stable when h increases from 0.8 to 2.4. In particular, when the predictive steps are long enough (after 500 steps), the predictive error of previous methods increases much faster than that of the EKATP.

Figures 4A,C,E shows that the predictive error of the EKATP remains stable when the noise intensity σ increases from 0 to 0.08 under complexity $h = 0.8$. **Figures 4B,D,F** shows that the predictive error of the EKATP remains stable when the noise intensity σ increases from 0 to 0.08 under complexity $h = 2.4$.

Figure 4 demonstrates that the predictive accuracy and robustness of the EKATP outperforms the existing methods under clean and noisy environments.

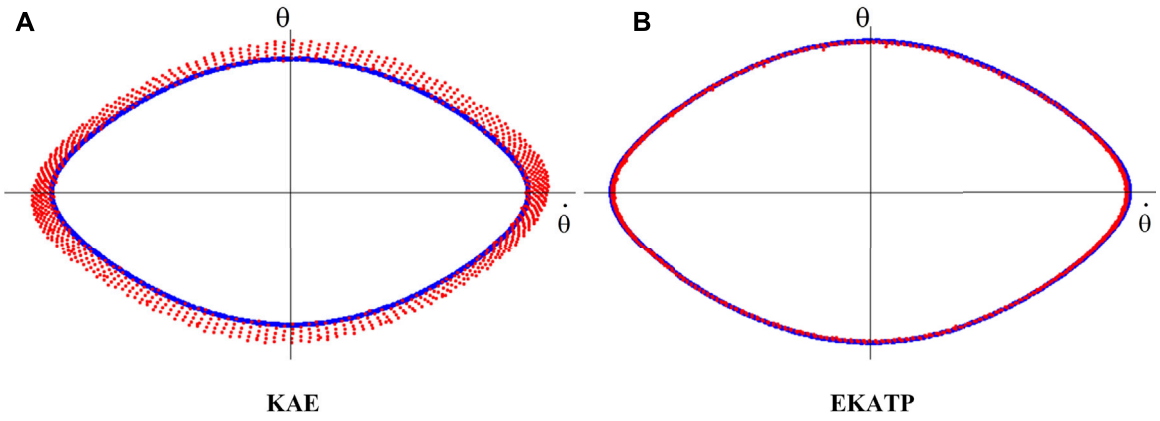

FIGURE 5 | The prediction trajectories within 1,000 steps under the initial conditions of $h = 2.4$ and $\sigma = 0.03$. The abscissa represents $\dot{\theta}$, the ordinate represents θ, blue colours represent true data and red dots represent predictive data. **(A)** KAE. **(B)** EKATP.

TABLE 1 | Predictive error at 1,000 steps for both the KAE and EKATP.

Model	h	θ	Predictive error				*p*-Value
			Min	Max	Avg	Var	
KAE	0.8	0.00	0.427	0.012	0.052	8.29e−03	2.60e−02
EKATP	0.8	0.00	**0.001**	**0.006**	**0.003**	**2.18e−06**	
KAE	0.8	0.03	0.038	0.253	0.112	2.80e−03	2.11e−04
EKATP	0.8	0.03	0.038	**0.089**	**0.058**	**1.42e−04**	
KAE	2.4	0.00	0.020	0.225	0.067	2.03e−03	5.79e−06
EKATP	2.4	0.00	**0.003**	**0.010**	**0.005**	**4.20e−06**	
KAE	2.4	0.03	0.030	0.967	0.131	4.04e−02	2.42e−02
EKATP	2.4	0.03	**0.011**	**0.040**	**0.021**	**5.79e−05**	

Since **Figure 4** shows that KAE has a better predictive effect than the other existing methods, we use it to compare the predictive performance with the EKATP by visualizing the predictive trajectory.

Indicated by our data preprocessing procedure, since the 2-dimensional time series state and 64-dimensional time series state are diffeomorphic (Sauer et al., 1991), the mapping between these two time series is reversible. Here, we map the 64-dimensional protein time series predictive results onto a 2-dimensional space by orthogonal inverse transformation (Anderson et al., 1987) to visualize the predictive time series trajectory. **Figure 5** shows the predictive trajectories of the KAE and EKATP within 1,000 steps under the initial conditions of $h = 2.4$ and $\sigma = 0.03$, which show that the predictive protein time series trajectory of the EKATP (**Figure 5B**) is much closer to the true trajectory than that of the KAE (**Figure 5A**). **Figure 5** further indicates that the predictive accuracy and robustness of the EKATP is better than that of the KAE.

To further prove that the EKATP has strong generalizability, we randomly selected 20 pieces of different protein time series data for model training and analysis. The details are listed in **Table 1**; **Supplementary Table S2.7**.

After we employ 20 different proteomics time series datasets to test the KAE and EKATP, **Table 1** shows the predictive error of the KAE and EKATP at 1,000 steps under different initial noise and complexity (h) conditions, which demonstrates that the EKATP has less of a statistically significant minimum, maximum, average and variance of the predictive error than the KAE under each noise and complexity (h) condition (p-value <0.05) (Gao et al., 2017; Li et al., 2017; Gao et al., 2021). **Table 1** implies that the EKATP has statistically significant predictive power for different time series datasets.

Metabolomics

We usually employ a nonlinear biological fluid system (Noack et al., 2003) to describe the high-dimensional metabolic time series with a low-dimensional manifold (Sauer et al., 1991) for the flow behavior of biological fluids simulation by **Eq. 21**,

$$\begin{cases} \dot{x} = \gamma x - \omega y + Axz \\ \dot{y} = \omega x + \gamma y + Ayz \\ \dot{z} = -\lambda\left(z - x^2 - y^2\right), \end{cases} \tag{21}$$

where γ, ω and A determine the size of the fluid. λ determines the speed of the dynamics of z. The different initial values of x, y and z determine the different nonlinear complexities of the metabolomics time series. We use the initial conditions ζ_1 (x=0, y = -0.01, z = 0) and ζ_2 (x=0.01, y = -0.1, z = 0.5) to generate a high-dimensional metabolomics time series with low complexity h_1 and high complexity h_2, respectively.

Since the metabolomics time series contains considerable noise, we employ white Gaussian noise (Li et al., 2017) to describe it by **Eq. 22**,

$$\begin{cases} \tilde{x} = x + \varepsilon_x \\ \tilde{y} = y + \varepsilon_y \\ \tilde{z} = z + \varepsilon_z \end{cases}. \tag{22}$$

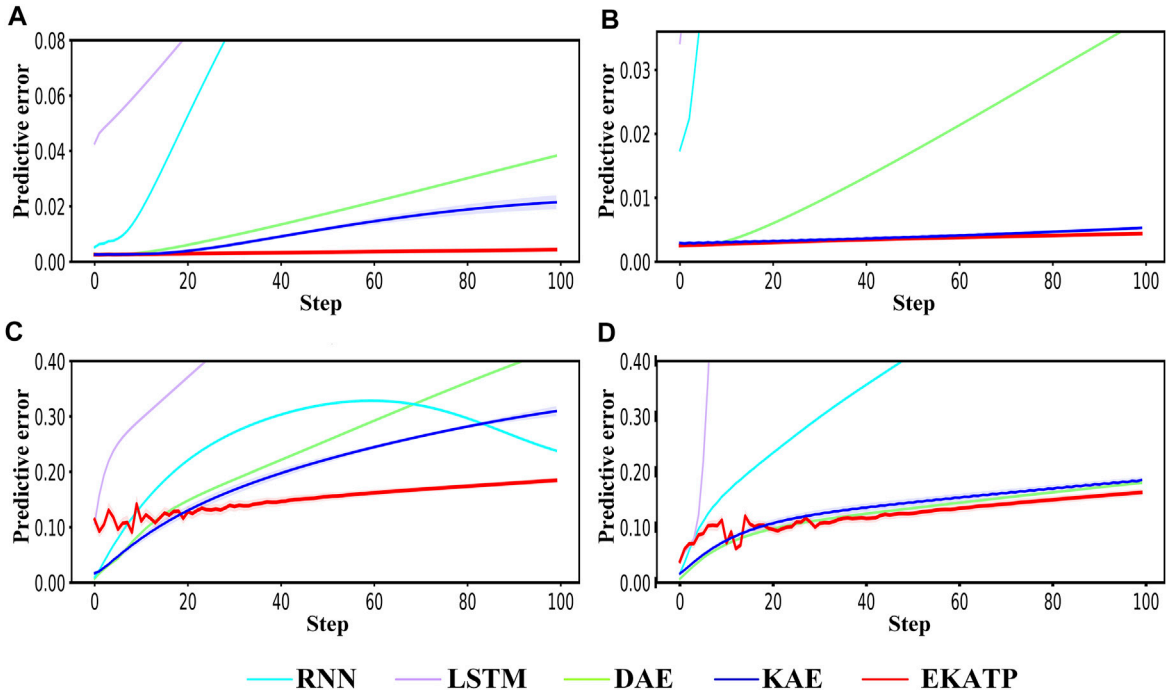

FIGURE 6 | Comparison of the RNN, LSTM, DAE, KAE and EKATP. The abscissa represents the time step, and the ordinate represents the predictive error. **(A)** The initial conditions are h_1 and $\sigma = 0.000$. **(B)** The initial conditions are h_1 and $\sigma = 0.001$. **(C)** The initial conditions are h_2 and $\sigma = 0.000$. **(D)** The initial conditions are h_2 and $\sigma = 0.001$.

where $\tilde{x}$, $\tilde{y}$ and $\tilde{z}$ represent data with noise. ε_x, ε_y and ε_z represent the white Gaussian noise for x, y and z by normal distributions $\mathcal{N}(0, \sigma^2)$ with a zero mean and a standard deviation σ.

Fig. 6 Comparison of the RNN, LSTM, DAE, KAE and EKATP. The abscissa represents the time step, and the ordinate represents the predictive error. (A) The initial conditions are h_1 and $\sigma = 0.000$. (B) The initial conditions are h_1 and $\sigma = 0.001$. (C) The initial conditions are h_2 and $\sigma = 0.000$. (D) The initial conditions are h_2 and $\sigma = 0.001$.

Here, we show how to generate a high-dimensional metabolomics time series with a low-dimensional manifold. First, we build up the 3-dimensional time series $\boldsymbol{V} = (\boldsymbol{V^1}, \boldsymbol{V^2}, \ldots, \boldsymbol{V^T}) \in \mathbb{R}^3$, which is listed in **Supplementary Tables S3.1; S3.2**. Next, we develop a random orthogonal transformation (Anderson et al., 1987) matrix $\boldsymbol{P} \in \mathbb{R}^{96\times3}$. Finally, we map the state of the 3-dimensional time series onto the state of the 96-dimensional time series by **Eq. 18** to simulate a high-dimensional metabolic time series $\boldsymbol{F} = (\boldsymbol{F^1}, \boldsymbol{F^2}, \ldots, \boldsymbol{F^T}) \in \mathbb{R}^{96}$ with the 3-dimensional manifold, which is listed in **Supplementary Tables S3.3, S3.4**.

To demonstrate the accuracy and robustness of the EKATP, we generate a system containing $T = 900$ steps and choose the last 100 steps to visualize the predictive result of the EKATP. **Figure 6** shows the predictive results of the metabolic time series under different initial conditions and environments for the last 100 steps. Detailed information is listed in **Supplementary Tables S3.5, S3.6**; **Supplementary Presentation S4**.

Figures 6A,C demonstrates that the EKATP has less of a predictive error under a clean environment (σ=0.000) than the existing methods. When the complexity of h increases, the predictive error of the EKATP remains stable. With the increase in the predictive step, the predictive error of the existing methods increases rapidly, while the predictive error of the EKATP remains small.

Figures 6B,D suggests that the EKATP not only has less of a predictive error under a low noise intensity environment (σ=0.001) than the existing methods but also has a predictive error of the EKATP that remains stable when h increases. In particular, when the predictive steps are long enough, the predictive error of the EKATP increases much slower than that of the existing methods.

Figure 6 implies that the EKATP has better predictive accuracy and robustness than the existing methods under clean and weakly noisy environments.

Since a metabolomics time series usually has strong noise intensity (Mak et al., 2015), we use the EKATP to predict a high-dimensional metabolomics time series under strong noise intensities to prove its robustness. Because the 3-dimensional time series state and the 96-dimensional time series state are diffeomorphic (Sauer et al., 1991), the mapping between these two time series is reversible. Thus, after we map the 96-dimensional metabolic time series predictive results onto a 3-dimensional space by orthogonal inverse transformation (Anderson et al., 1987), **Figure 7** shows the predictive time series trajectories by the EKATP under different intensities of noise. We select the last 100

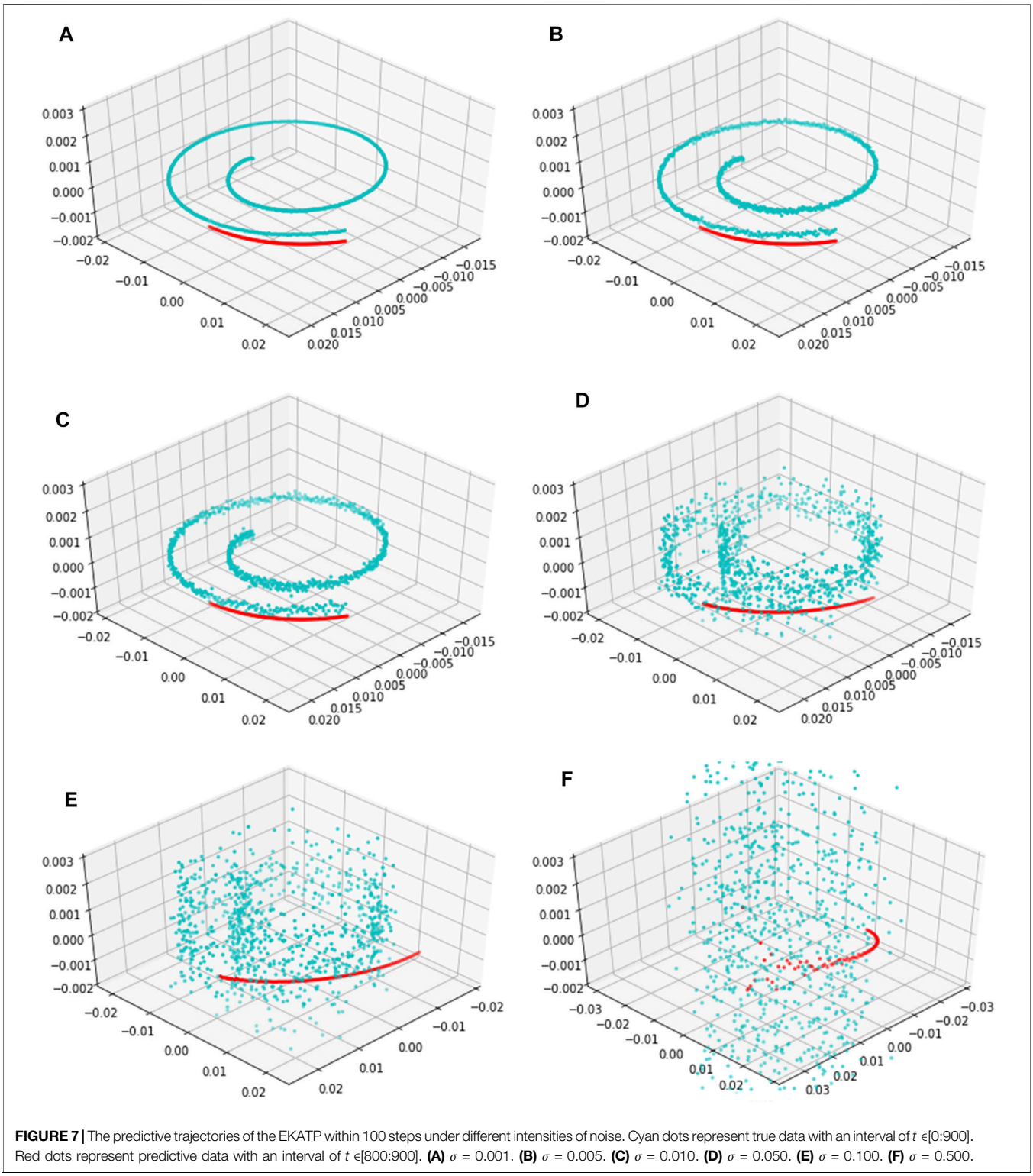

FIGURE 7 | The predictive trajectories of the EKATP within 100 steps under different intensities of noise. Cyan dots represent true data with an interval of $t \in$[0:900]. Red dots represent predictive data with an interval of $t \in$[800:900]. **(A)** σ = 0.001. **(B)** σ = 0.005. **(C)** σ = 0.010. **(D)** σ = 0.050. **(E)** σ = 0.100. **(F)** σ = 0.500.

steps to validate the predictive power of the EKATP as in the previous setup (**Supplementary Table S3.6**). The results demonstrate that although the true data become gradually messy when we increase the noise intensity σ, the predictive time series trajectory of the EKATP is still very close to the true data to a certain extent (**Figures 7A,B,C,D,E,F**), which implies that the EKATP still has a satisfactory predictive performance when we increase the noise intensity.

Moreover, we use **Eqs 23**, **24** to calculate the Pearson correlation coefficient (PCC) (Abar et al., 2017) and the root

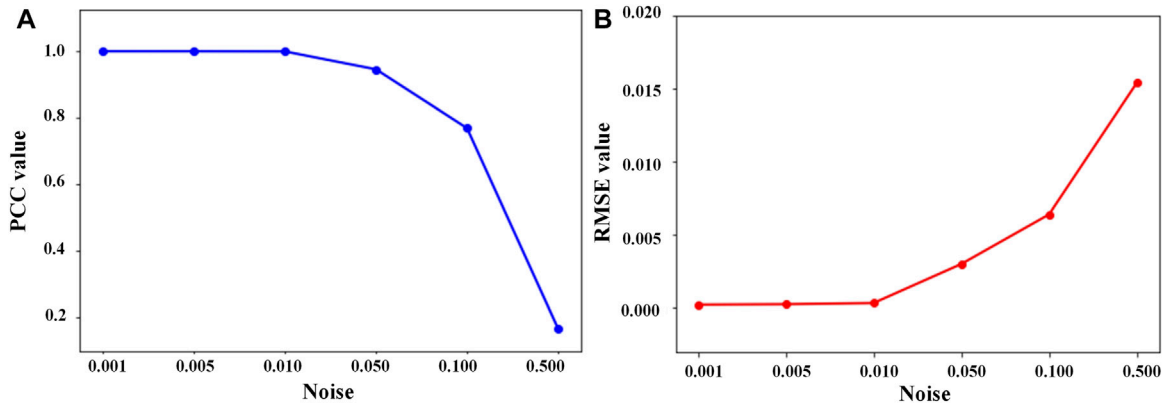

FIGURE 8 | PCC and RMSE values between predictive and true data under different noise intensities. The details are listed in **Supplementary Table S3.7**. **(A)** PCC value between predictive and true data, where the abscissa represents the noise intensity and the ordinate represents the PCC value. **(B)** RMSE value between predictive and true data, where the abscissa represents the noise intensity and the ordinate represents the RMSE value.

mean squared error (RMSE) (Abar et al., 2017) between predictive and true data under different noise intensities.

Here, $\boldsymbol{V^t}$ and $\hat{\boldsymbol{V}}^{\boldsymbol{t}}$ represent the true and predictive data at time t. $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ represent the average value for true and predictive data, respectively. p represents the predictive step size.

$$\mathrm{PCC} = \frac{\sum_{t=1}^{t=p}\left(\hat{\boldsymbol{V}}^{\boldsymbol{t}} - \hat{\boldsymbol{\mu}}\right)\left(\boldsymbol{V^t} - \boldsymbol{\mu}\right)}{\sqrt{\sum_{t=1}^{t=p}\left(\hat{\boldsymbol{V}}^{\boldsymbol{t}} - \hat{\boldsymbol{\mu}}\right)^2}\sqrt{\sum_{t=1}^{t=p}\left(\boldsymbol{V^t} - \boldsymbol{\mu}\right)^2}}. \tag{23}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{p}\sum_{t=1}^{t=p}\left\|\hat{\boldsymbol{V}}^{\boldsymbol{t}} - \boldsymbol{V^t}\right\|^2}. \tag{24}$$

Figure 8A shows that the PCC value of the EKATP gradually decreases when we increase the noise intensity σ, but the overall value is relatively high. **Figure 8B** indicates that with the increase in noise intensity σ, although the RMSE value of the EKATP gradually increases, it is still relatively small. Thus, we conclude that the EKATP can effectively avoid noise interference and is robust enough under a very strong noise intensity condition.

CONCLUSION AND FUTURE WORK

To answer the three proposed questions, this study developed an EKATP to predict the future state of a high-dimensional nonlinear multi-omics time series. First, we select key features from high-dimensional nonlinear multi-omics time series data. After that, we map these key features to the low-dimensional linear space. Next, we obtain the future state of the multi-omics time series by learning the evolutionary relationship between the adjacent states of the time series in the low-dimensional linear space. Finally, we predict the future state of the high-dimensional nonlinear multi-omics time series by mapping the low-dimensional linear predictive state back to the high-dimensional nonlinear space. The experimental results demonstrate that the EKATP can greatly improve the accuracy, robustness and generalisability to predict the future state of a time series for genomics (**Figures 2**, **3**), proteomics (**Figures 4**, **5**; **Table 1**) and metabolomics (**Figures 6–8**) datasets.

However, there are still several shortcomings to the current study. For example, we are still unclear on the impact of embedding dimensions from high-dimensional nonlinear space to low-dimensional linear space on predictive accuracy and the way to use high-performance computing to increase the efficiency of the EKATP. Applying the EKATP to network biological datasets (Liu X. et al., 2020) is also the direction we need to continue the study. Thus, we will improve the EKATP from these perspectives in the distant future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

YY, LZ and SL conceived the project. SL, YY and ZT carried out experiments. SL and YY visualized experiment results. SL drafted the manuscript. YY and LZ revised the article. All the authors read and approved the final article.

FUNDING

This work was supported by the National Science and Technology Major Project (2018ZX10201002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.761629/full#supplementary-material

REFERENCES

Abar, T., El Asmi, A. S., and Asmi, S. E. (2017). "Machine Learning Based QoE Prediction in SDN Networks," in Proceedings of the 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, Spain, June 2017. doi:10.1109/IWCMC.2017.7986488

Anderson, T. W., Olkin, I., and Underhill, L. G. (1987). Generation of Random Orthogonal Matrices. *SIAM J. Sci. Stat. Comput.* 8 (4), 625–629. doi:10.1137/0908055

Azencot, O., Erichson, N. B., Lin, V., and Mahoney, M. (2020). "Forecasting Sequential Data Using Consistent Koopman Autoencoders," in Proceedings of the 37th International Conference on Machine Learning, 475–485.

Bianconi, F., Antonini, C., Tomassoni, L., and Valigi, P. (2020). Robust Calibration of High Dimension Nonlinear Dynamical Models for Omics Data: An Application in Cancer Systems Biology. *IEEE Trans. Contr. Syst. Technol.* 28 (1), 196–207. doi:10.1109/TCST.2018.2844362

Chen, P., Liu, R., Aihara, K., and Chen, L. (2020). Autoreservoir Computing for Multistep Ahead Prediction Based on the Spatiotemporal Information Transformation. *Nat. Commun.* 11 (1), 4568. doi:10.1038/s41467-020-18381-0

Davidson, E., and Levin, M. (2005). Gene Regulatory Networks. *Proc. Natl. Acad. Sci.* 102 (14), 4935. doi:10.1073/pnas.0502024102

Eisenhammer, T., Hübler, A., Packard, N., and Kelso, J. A. S. (1991). Modeling Experimental Time Series with Ordinary Differential Equations. *Biol. Cybern.* 65 (2), 107–112. doi:10.1007/BF00202385

Fischer, H. P. (2008). Mathematical Modeling of Complex Biological Systems: from Parts Lists to Understanding Systems Behavior. *Alcohol. Res. Health* 31 (1), 49–59.

Gao, H., Yin, Z., Cao, Z., and Zhang, L. (2017). Developing an Agent-Based Drug Model to Investigate the Synergistic Effects of Drug Combinations. *Molecules* 22 (12), 2209. doi:10.3390/molecules22122209

Gao, J., Liu, P., Liu, G.-D., and Zhang, L. (2021). Robust Needle Localization and Enhancement Algorithm for Ultrasound by Deep Learning and Beam Steering Methods. *J. Comput. Sci. Technol.* 36 (2), 334–346. doi:10.1007/s11390-021-0861-7

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science* 313 (5786), 504–507. doi:10.1126/science.1127647

Hirsch, M. (1974). *Differential Equations, Dynamical Systems, and Linear Algebra Mathematics*. America: Academic press. doi:10.1016/s0079-8169(08)x6044-1

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Holmes, P., Lumley, J. L., Berkooz, G., and Rowley, C. W. (2012). *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge, UK, Cambridge University Press.

Iuchi, H., Sugimoto, M., and Tomita, M. (2018). MICOP: Maximal Information Coefficient-Based Oscillation Prediction to Detect Biological Rhythms in Proteomics Data. *BMC Bioinformatics* 19 (1), 249. doi:10.1186/s12859-018-2257-4

Ji, Z., Yan, K., Li, W., Hu, H., and Zhu, X. (2017). Mathematical and Computational Modeling in Complex Biological Systems. *Biomed. Res. Int.* 2017, 1–16. doi:10.1155/2017/5958321

Jiang, J., and Lai, Y.-C. (2019). Model-free Prediction of Spatiotemporal Dynamical Systems with Recurrent Neural Networks: Role of Network Spectral Radius. *Phys. Rev. Res.* 1 (3), 033056. doi:10.1103/PhysRevResearch.1.033056

Koopman, B. O. (1931). Hamiltonian Systems and Transformation in Hilbert Space. *Proc. Natl. Acad. Sci.* 17 (5), 315–318. doi:10.1073/pnas.17.5.315

Lai, Q., Zhao, X.-W., Huang, J.-N., Pham, V.-T., and Rajagopal, K. (2018). Monostability, Bistability, Periodicity and Chaos in Gene Regulatory Network. *Eur. Phys. J. Spec. Top.* 227 (7), 719–730. doi:10.1140/epjst/e2018-700132-8

Levnajić, Z., and Tadić, B. (2010). Stability and Chaos in Coupled Two-Dimensional Maps on Gene Regulatory Network of Bacterium *E. coli*. *Chaos* 20 (3), 033115. doi:10.1063/1.3474906

Li, T., Cheng, Z., and Zhang, L. (2017). Developing a Novel Parameter Estimation Method for Agent-Based Model in Immune System Simulation under the Framework of History Matching: A Case Study on Influenza A Virus Infection. *Ijms* 18 (12), 2592. doi:10.3390/ijms18122592

Liang, Y., and Kelemen, A. (2017a). Computational Dynamic Approaches for Temporal Omics Data with Applications to Systems Medicine. *BioData Mining* 10 (1), 20. doi:10.1186/s13040-017-0140-x

Liang, Y., and Kelemen, A. (2017b). Dynamic Modeling and Network Approaches for Omics Time Course Data: Overview of Computational Approaches and Applications. *Brief. Bioinform.* 19 (5), 1051–1068. doi:10.1093/bib/bbx036

Liu, G.-D., Li, Y.-C., Zhang, W., and Zhang, L. (2020). A Brief Review of Artificial Intelligence Applications and Algorithms for Psychiatric Disorders. *Engineering* 6 (4), 462–467. doi:10.1016/j.eng.2019.06.008

Liu, X., Maiorino, E., Halu, A., Glass, K., Prasad, R. B., Loscalzo, J., et al. (2020). Robustness and Lethality in Multilayer Biological Molecular Networks. *Nat. Commun.* 11 (1), 6043. doi:10.1038/s41467-020-19841-3

Lockhart, D. J., and Winzeler, E. A. (2000). Genomics, Gene Expression and DNA Arrays. *Nature* 405 (6788), 827–836. doi:10.1038/35015701

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *J. Atmos. Sci.* 20 (2), 130–141. doi:10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2

Lusch, B., Kutz, J. N., and Brunton, S. L. (2018). Deep Learning for Universal Linear Embeddings of Nonlinear Dynamics. *Nat. Commun.* 9 (1), 4950. doi:10.1038/s41467-018-07210-0

Mak, T. D., Laiakis, E. C., Goudarzi, M., and Fornace, A. J., Jr. (2015). Selective Paired Ion Contrast Analysis: a Novel Algorithm for Analyzing Postprocessed LC-MS Metabolomics Data Possessing High Experimental Noise. *Anal. Chem.* 87 (6), 3177–3186. doi:10.1021/ac504012a

Mann, M., Hendrickson, R. C., and Pandey, A. (2001). Analysis of Proteins and Proteomes by Mass Spectrometry. *Annu. Rev. Biochem.* 70 (1), 437–473. doi:10.1146/annurev.biochem.70.1.437

Noack, B. R., Afanasiev, K., Morzyński, M., Tadmor, G., and Thiele, F. (2003). A Hierarchy of Low-Dimensional Models for the Transient and post-transient cylinder Wake. *J. Fluid Mech.* 497, 335–363. doi:10.1017/S0022112003006694

Perez-Riverol, Y., Kuhn, M., Vizcaíno, J. A., Hitz, M.-P., and Audain, E. (2017). Accurate and Fast Feature Selection Workflow for High-Dimensional Omics Data. *PloS one* 12 (12), e0189875. doi:10.1371/journal.pone.0189875

Sauer, T., Yorke, J. A., and Casdagli, M. (1991). Embedology. *J. Stat. Phys.* 65 (3), 579–616. doi:10.1007/BF01053745

Sevim, V., and Rikvold, P. A. (2008). Chaotic Gene Regulatory Networks Can Be Robust against Mutations and Noise. *J. Theor. Biol.* 253 (2), 323–332. doi:10.1016/j.jtbi.2008.03.003

Song, H., Jiang, Z., Men, A., and Yang, B. (2017). A Hybrid Semi-supervised Anomaly Detection Model for High-Dimensional Data. *Comput. Intelligence Neurosci.* 2017, 1–9. doi:10.1155/2017/8501683

Soon, W. W., Hariharan, M., and Snyder, M. P. (2013). High-throughput Sequencing for Biology and Medicine. *Mol. Syst. Biol.* 9 (1), 640. doi:10.1038/msb.2012.61

Suzuki, Y., Lu, M., Ben-Jacob, E., and Onuchic, J. N. (2016). Periodic, Quasi-Periodic and Chaotic Dynamics in Simple Gene Elements with Time Delays. *Sci. Rep.* 6 (1), 21037. doi:10.1038/srep21037

Tsimring, L. S. (2014). Noise in Biology. *Rep. Prog. Phys.* 77 (2)–026601. doi:10.1088/0034-4885/77/2/026601

Tyers, M., and Mann, M. (2003). From Genomics to Proteomics. *Nature* 422 (6928), 193–197. doi:10.1038/nature01510

Wang, H., Li, M., and Yue, X. (2021). IncLSTM: Incremental Ensemble LSTM Model towards Time Series Data. *Comput. Electr. Eng.* 92, 107156. doi:10.1016/j.compeleceng.2021.107156

Wang, W., Huang, Y., Wang, Y., and Wang, L. (2014). "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction," in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, June 2014. doi:10.1109/CVPRW.2014.79

Weckwerth, W. (2003). Metabolomics in Systems Biology. *Annu. Rev. Plant Biol.* 54 (1), 669–689. doi:10.1146/annurev.arplant.54.031902.135014

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. (2009). Integrated Network Analysis Platform for Protein-Protein Interactions. *Nat. Methods* 6 (1), 75–77. doi:10.1038/nmeth.1282

Wu, W., Song, L., Yang, Y., Wang, J., Liu, H., and Zhang, L. (2020). Exploring the Dynamics and Interplay of Human Papillomavirus and Cervical Tumorigenesis by Integrating Biological Data into a Mathematical Model. *BMC Bioinformatics* 21 (7), 152. doi:10.1186/s12859-020-3454-5

Xia, Y., Yang, C., Hu, N., Yang, Z., He, X., Li, T., et al. (2017). Exploring the Key Genes and Signaling Transduction Pathways Related to the Survival Time of Glioblastoma Multiforme Patients by a Novel Survival Analysis Model. *BMC Genomics* 18 (1), 950. doi:10.1186/s12864-016-3256-3

Xiao, M., Liu, G., Xie, J., Dai, Z., Wei, Z., Ren, Z., et al. (2021). 2019nCoVAS: Developing the Web Service for Epidemic Transmission Prediction, Genome Analysis, and Psychological Stress Assessment for 2019-nCoV. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18 (4), 1250–1261. doi:10.1109/TCBB.2021.3049617

Xiao, M., Yang, X., Yu, J., and Zhang, L. (2020). CGIDLA:Developing the Web Server for CpG Island Related Density and LAUPs (Lineage-Associated Underrepresented Permutations) Study. *Ieee/acm Trans. Comput. Biol. Bioinf.* 17 (6), 2148–2154. doi:10.1109/TCBB.2019.2935971

You, Y., Ru, X., Lei, W., Li, T., Xiao, M., Zheng, H., et al. (2020). Developing the Novel Bioinformatics Algorithms to Systematically Investigate the Connections Among Survival Time, Key Genes and Proteins for Glioblastoma Multiforme. *BMC Bioinformatics* 21 (13), 383. doi:10.1186/s12859-020-03674-4

Zhang, A., Sun, H., Wang, P., Han, Y., and Wang, X. (2012a). Recent and Potential Developments of Biofluid Analyses in Metabolomics. *J. Proteomics* 75 (4), 1079–1088. doi:10.1016/j.jprot.2011.10.027

Zhang, Z., Ye, W., Qian, Y., Zheng, Z., Huang, X., and Hu, G. (2012b). Chaotic Motifs in Gene Regulatory Networks. *PLOS ONE* 7 (7), e39355. doi:10.1371/journal.pone.0039355

Zhang, L., Dai, Z., Yu, J., and Xiao, M. (2020). CpG-island-based Annotation and Analysis of Human Housekeeping Genes. *Brief. Bioinform.* 22 (1), 515–525. doi:10.1093/bib/bbz134

Zhang, L., Fu, C., Li, J., Zhao, Z., Hou, Y., Zhou, W., et al. (2019a). Discovery of a Ruthenium Complex for the Theranosis of Glioma through Targeting the Mitochondrial DNA with Bioinformatic Methods. *Ijms* 20 (18), 4643. doi:10.3390/ijms20184643

Zhang, L., Li, J., Yin, K., Jiang, Z., Li, T., Hu, R., et al. (2019b). Computed Tomography Angiography-Based Analysis of High-Risk Intracerebral Haemorrhage Patients by Employing a Mathematical Model. *BMC Bioinformatics* 20 (7), 193. doi:10.1186/s12859-019-2741-5

Zhang, L., Liu, G., Kong, M., Li, T., Wu, D., Zhou, X., et al. (2019c). Revealing Dynamic Regulations and the Related Key Proteins of Myeloma-Initiating Cells by Integrating Experimental Data into a Systems Biological Model. *Bioinformatics* 37 (11), 1554–1561. doi:10.1093/bioinformatics/btz542

Zhang, L., Qiao, M., Gao, H., Hu, B., Tan, H., Zhou, X., et al. (2016). Investigation of Mechanism of Bone Regeneration in a Porous Biodegradable Calcium Phosphate (CaP) Scaffold by a Combination of a Multi-Scale Agent-Based Model and Experimental Optimization/validation. *Nanoscale* 8 (31), 14877–14887. doi:10.1039/C6NR01637E

Zhang, L., Xiao, M., Zhou, J., and Yu, J. (2018). Lineage-associated Underrepresented Permutations (LAUPs) of Mammalian Genomic Sequences Based on a Jellyfish-Based LAUPs Analysis Application (JBLA). *Bioinformatics* 34 (21), 3624–3630. doi:10.1093/bioinformatics/bty392

Zhang, L., Zhang, L., Guo, Y., Xiao, M., Feng, L., Yang, C., et al. (2021a). MCDB: A Comprehensive Curated Mitotic Catastrophe Database for Retrieval, Protein Sequence Alignment, and Target Prediction. *Acta Pharmaceutica Sinica B*. doi:10.1016/j.apsb.2021.05.032

Zhang, L., and Zhang, S. (2017). Using Game Theory to Investigate the Epigenetic Control Mechanisms of Embryo Development. *Phys. Life Rev.* 20, 140–142. doi:10.1016/j.plrev.2017.01.007

Zhang, L., Zhao, J., Bi, H., Yang, X., Zhang, Z., Su, Y., et al. (2021b). Bioinformatic Analysis of Chromatin Organization and Biased Expression of Duplicated Genes between Two Poplars with a Common Whole-Genome Duplication. *Hortic. Res.* 8 (1), 62. doi:10.1038/s41438-021-00494-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 24 November 2021
doi: 10.3389/fgene.2021.729326

Credible Mendelian Randomization Studies in the Presence of Selection Bias Using Control Exposures

Zhao Yang[1], *C. Mary Schooling*[1,2] *and Man Ki Kwok*[1]*

[1]School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China, [2]Graduate School of Public Health and Health Policy, City University of New York, New York, NY, United States

Selection bias is increasingly acknowledged as a limitation of Mendelian randomization (MR). However, few methods exist to assess this issue. We focus on two plausible causal structures relevant to MR studies and illustrate the data-generating process underlying selection bias *via* simulation studies. We conceptualize the use of control exposures to validate MR estimates derived from selected samples by detecting potential selection bias and reproducing the exposure–outcome association of primary interest based on subject matter knowledge. We discuss the criteria for choosing the control exposures. We apply the proposal in an MR study investigating the potential effect of higher transferrin with stroke (including ischemic and cardioembolic stroke) using transferrin saturation and iron status as control exposures. Theoretically, selection bias affects associations of genetic instruments with the outcome in selected samples, violating the exclusion-restriction assumption and distorting MR estimates. Our applied example showing inconsistent effects of genetically predicted higher transferrin and higher transferrin saturation on stroke suggests the potential selection bias. Furthermore, the expected associations of genetically predicted higher iron status on stroke and longevity indicate no systematic selection bias. The routine use of control exposures in MR studies provides a valuable tool to validate estimated causal effects. Like the applied example, an antagonist, decoy, or exposure with similar biological activity as the exposure of primary interest, which has the same potential selection bias sources as the exposure–outcome association, is suggested as the control exposure. An additional or a validated control exposure with a well-established association with the outcome is also recommended to explore possible systematic selection bias.

Keywords: causal estimates, control exposures, Mendelian randomization, reproducible, selection bias

OPEN ACCESS

Edited by:
Miguel E. Rentería, QIMR Berghofer Medical Research Institute, Australia

Reviewed by:
Huaizhen Qin, University of Florida, United States
Santiago Diaz-Torres, The University of Queensland, Australia

***Correspondence:**
Man Ki Kwok
maggiek@hku.hk

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *23 June 2021*
Accepted: *01 November 2021*
Published: *24 November 2021*

Citation:
Yang Z, Schooling CM and Kwok MK (2021) Credible Mendelian Randomization Studies in the Presence of Selection Bias Using Control Exposures. Front. Genet. 12:729326. doi: 10.3389/fgene.2021.729326

HIGHLIGHTS

What is Already Known on this Subject?

- Mendelian randomization (MR) provides unconfounded estimates, but is particularly vulnerable to selection bias because of the small magnitude of genetic estimates.

Abbreviations: CR, competing risks; DAG, directed acyclic graph; GWAS, genome-wide association study; IV, instrumental variable; MR, Mendelian randomization.

- Negative controls provide helpful tools to detect residual confounding, selection, and measurement bias in conventional epidemiological studies but often lack specificity in the type of bias they detect.

What this Adds to What is Known?

- Given genetics are a lifelong exposure, a key source of selection bias in MR studies is missing people from the same underlying birth cohorts as the original population who die before recruitment, which may violate the exclusion-restriction assumption and distort the MR estimates.
- The use of control exposures that have the same potential selection bias sources as the exposure–outcome association of interest can detect potential selection bias and validate MR estimates.
- The estimated exposure–outcome association is more credible if this result is robust to potential selection bias and reproducible by using the relevant control exposures based on subject matter knowledge.

What is the Implication, What Should Change Now?

- Systematic selection bias may occur particularly when the genetic variants affect survival and the outcome of interest or a competing risk of that outcome affects survival; interpretation of MR estimates should be cautious.
- The routine use of control exposures could add more credibility to MR estimates.

INTRODUCTION

Mendelian randomization (MR) uses genetic variants as a natural experiment in observational studies to investigate potential causal effects of modifiable risk factors on health outcomes (Davey Smith and Ebrahim, 2003). MR is often conducted in two homogeneous study populations, i.e., two-sample MR (Burgess et al., 2015). MR is thought to be robust to the confounding that often occurs in conventional observational studies due to the random allocation of genetic endowment at conception being used as a proxy for the exposure (Burgess et al., 2012; Davies et al., 2018). Currently, MR is a popular approach for assessing causality (Sekula et al., 2016). However, MR estimate rests on stringent assumptions, as illustrated using directed acyclic graphs (DAGs) in **Figures 1A,B** (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008).

- IV1 (the relevance assumption): the genetic variant is robustly associated with the exposure of interest;
- IV2 (the independence assumption): the genetic variant is not associated with confounders that bias the exposure–outcome association;
- IV3 (the exclusion-restriction assumption): the genetic variant affects the health outcome only *via* its effect on the exposure.

Notably, aside from IV1 that can be empirically verified using the F-statistic (Staiger and Stock, 1997; Bowden et al., 2016a), IV2 and IV3 are typically harder to justify. Hence, violations of these assumptions can occur, leading to misleading conclusions. Of these, selection bias is increasingly acknowledged as distorting MR estimates in the selected populations investigated (Nitsch et al., 2006; VanderWeele et al., 2014; Canan et al., 2017; Vansteelandt et al., 2018a; Vansteelandt et al., 2018b; Munafò et al., 2018; Munafò and Smith, 2018; Gkatzionis and Burgess, 2019; Smit et al., 2019; Schooling et al., 2020) and has largely focused on bias arising from selection on exposure (Vansteelandt et al., 2018a; Vansteelandt et al., 2018b; Munafò et al., 2018; Gkatzionis and Burgess, 2019; Smit et al., 2019).

Genetic studies are usually carefully designed to avoid selecting sample on genetic make-up and phenotypes. Generally, selection bias occurs in an MR study when the sample in the original genome-wide association study (GWAS) are selected conditional on survival until study recruitment on genotype of interest in the presence of prior death from the outcome or competing risks of the outcome (**Figure 1C**), especially in the original outcome GWAS (Schooling et al., 2020). The problem is the time lag between genetic randomization at conception and recruitment of participants into the GWAS. Participants diagnosed with or dead from the outcome or a competing risk of the outcome are not recruited into the outcome GWAS, which attenuates or reverses MR estimates for harmful exposures, because people who have already died of their harmful genetic endowment and people who have died of the outcome or a competing risk of the outcome are missing. As such, selection bias may create a spurious genetic variant–outcome association by opening the backdoor path from genetic instruments to the outcome of interest, violating the IV3 assumption.

For example, previous observational studies showed that higher transferrin binds to circulating iron and influences iron status, which may further cause iron-deficiency anemia and increase the risk of stroke (Chang et al., 2013; Marniemi et al., 2005; Gillum et al., 1996). However, a recent MR study reported that lower iron status also appeared to protect against stroke (van der et al., 2005; Gill et al., 2018), especially cardioembolic stroke (Gill et al., 2018). An increasingly acknowledged explanation is selection bias, possibly due to the presence of competing risks [e.g., coronary artery disease (Gill et al., 2017), hypercholesterolemia (Gill et al., 2019), chronic kidney disease (Fishbane et al., 2009), skin infections (Gill et al., 2019), liver disorders (e.g., hepatitis C) (Shan et al., 2005), and rheumatoid arthritis (Yuan and Larsson, 2020)] caused by the shared confounders (e.g., socioeconomic position, lifestyle, and health status), affecting survival of the underlying population (Camaschella, 2015; McLean et al., 2009), as shown in **Figure 2**. For instance, people with competing risks, such as coronary artery disease, tend to die earlier than those with stroke in Western settings (Kesteloot and Decramer, 2008; Menotti et al., 2019; Diseases and Injuries, 2020). As such, people vulnerable to these competing risks with higher iron status may die before

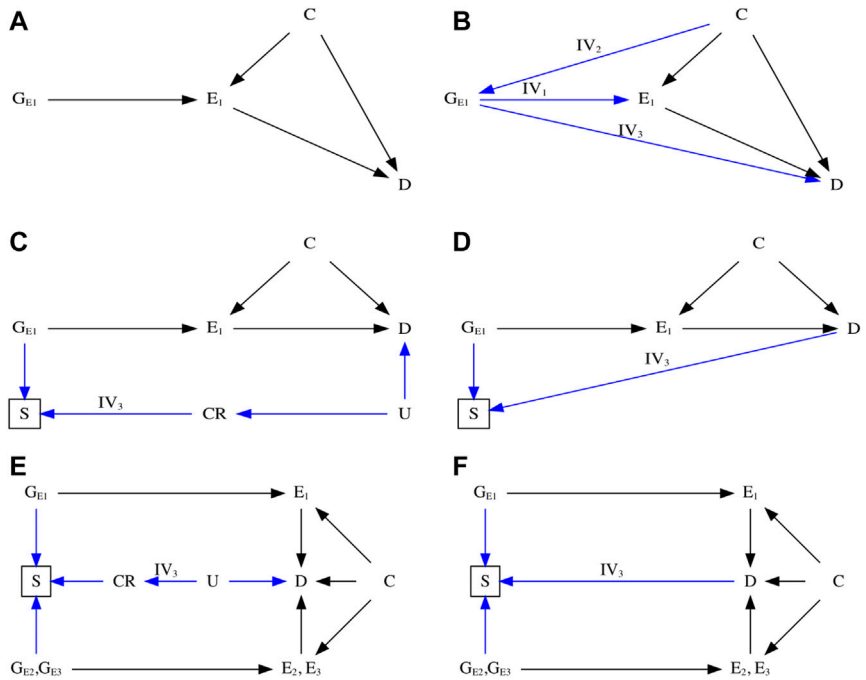

FIGURE 1 | Directed acyclic graph (DAG) illustrating Mendelian randomization (MR). **(A)** DAG illustrating an ideal scenario of an MR study. **(B)** DAG illustrating the three instrumental assumptions (Davey Smith and Ebrahim, 2003)—IV1: Relevance (Burgess et al., 2015); IV2: Independence (Burgess et al., 2012); IV3: Exclusion restriction. **(C)** DAG illustrating potential biased pathway with selection bias in the presence of competing risks that share substantial etiological factors with the outcome. **(D)** DAG illustrating potential biased pathway with selection bias in the unrepresentative selected samples. **(E)** DAG illustrating an MR study using control exposures to detect potential selection bias in the presence of competing risks. **(F)** DAG illustrating an MR study using control exposures to detect potential selection bias in the unrepresentative selected samples. E_1: the primary exposure of interest; E_2 and E_3: the control exposures; C: the confounder that associates with both the exposure and outcome; D: the outcome; CR: the competing risks; U: the unmeasured and shared confounders of the competing risks and the outcome; G_{E1}, G_{E2}, and G_{E3}: genetic variants that are strongly associated with the exposure of primary interest and the control exposures.

study recruitment, leaving more "healthier" participants in the study and inducing biased estimates.

Several statistical methods have been proposed to detect and eliminate selection bias in MR studies, most of which focus on bias arising from selection on exposure (Bareinboim and Pearl, 2012; Arnold and Ercumen, 2016; Hemani et al., 2017; Tchetgen Tchetgen and Wirth, 2017; Vansteelandt et al., 2018a; Brumpton et al., 2020; Zhao et al., 2020; Sanderson et al., 2021; Wang and Han, 2021), which is generally thought to have limited effects. However, selection on genetic endowment and outcome or competing risk of the outcome is more pervasive (Schooling et al., 2020) and can have larger effects. One approach that has not been considered is the use of a "negative control," which has been widely used in laboratory science for decades to help detect problems with the experimental method (Arnold and Ercumen, 2016). In epidemiological studies, a formal approach has been described in detail and suggested as a means of detecting residual confounding, selection bias, and measurement bias (Lipsitch et al., 2010; Arnold et al., 2016). Recently, negative control outcomes, defined as sharing identical confounders with the exposure–outcome association but not associated with the exposure, have been proposed to detect potential population stratification in MR studies (Sanderson et al., 2021). Other approaches include summary data-based MR [SMR, e.g., MR robust adjusted profile score (MR-RAPS)] (Zhao et al., 2020; Wang and Han, 2021), two-sample MR Steiger method (Hemani et al., 2017), and three-sample MR (Zhao et al., 2019), in which the selection procedure of genetic instrument (e.g., winner's curse) is considered a form of selection bias (Wang and Han, 2021). However, such a situation is different from the scenario where the original outcome GWAS is missing people from the

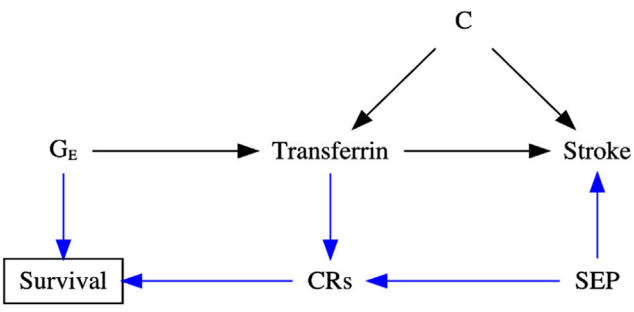

FIGURE 2 | Directed acyclic graph (DAG) illustrating the possible data-generating process underlying selection bias in the transferrin–stroke association due to missing people in the presence of competing risks (CRs, e.g., coronary artery disease) caused by the shared confounder [e.g., socioeconomic position (SEP)] of stroke and CRs in two-sample Mendelian randomization settings. C: the unmeasured confounder of the transferrin–stroke association.

same underlying population (birth cohorts) as those included, some of whom have already died from the instrument and some of whom have already died from the outcome or a competing risk of the outcome, as shown in **Figures 1C,D**.

In this study, as an extension of negative control outcomes, we advance the use of control exposures to validate MR estimates that might be susceptible to such selection bias. We focus on plausible causal structures relevant to MR studies and illustrate how to validate MR estimates using control exposures through a real example investigating the potential association of transferrin with stroke (including ischemic and cardioembolic stroke). This association is thought to be particularly vulnerable to selection bias, especially among older populations, because transferrin affects survival and stroke is open to competing risk from IHD (Schooling et al., 2020; Yang et al., 2021). We further discuss the criteria for choosing the control exposures and the limitations of this approach.

METHODS

Figures 1C,D show DAGs for MR with selection bias caused by sample selection. In the presence of competing risks (**Figure 1C**), the selected samples may have a lower risk of developing the phenotype [e.g., the outcome (D)] because the GWAS is missing people with genetic vulnerability to earlier death and people who have died from a disease that shares causes (e.g., U) with the phenotype. As such, the backdoor pathway directly linking G_{E1} to D will be reopened in the selected samples if the instruments affect survival, i.e., have allele frequencies that differ from the underlying population (e.g., birth cohort). This situation violates the IV3 assumption and distorts MR estimates, which can attenuate or reverse the true association or create a spurious association. The small effect sizes of genetic associations (Park et al., 2011; Global Burden of Disease, 2020) make them particularly vulnerable to perturbation by such bias (Schooling et al., 2020). In the absence of competing risks (**Figure 1D**), the phenotype (e.g., D) risk and instruments' frequencies may vary because of selecting on genetic instruments and outcome, which generates unrecoverable selection bias.

To clearly illustrate the data-generating process underlying selection bias due to missing people from the original birth cohorts who formed the underlying population through death before study recruitment, we conducted extensive simulation studies. Details are presented in the **Supplementary Material**. Briefly, we induced selection bias by selecting study participants as survivors to study recruitment. We assumed that the survival of the underlying population was influenced by the genetic instruments G_{E1}, exposure E_1, outcome D, confounder C of the exposure–outcome association, or the unmeasured confounder U mediated by competing risks CR. We used the relative hazard (i.e., hazard ratio) per-unit change in either G_{E1}, E_1, C, D, or U to quantify their effects on the survival, as shown in **Supplementary Figure S1**. As such, the impact of selection bias induced by the survival status of the underlying population until study recruitment was governed by hazard ratio of per-unit change in either G_{E1}, E_1, C, D, or U. Then, we induced selection bias in two-sample MR by having instruments determining survival to recruitment and outcome of interest affecting survival to recruitment. Details of the simulation study are in the **Supplementary Material**, along with the corresponding R scripts.

Figure 3 and **Supplementary Figure S1** show the impact of selection bias arising from selecting samples conditioning on genetic instruments G and outcome D, with no effects of either exposure E_1 or the shared confounder U mediated by competing risks on survival of the underlying population (i.e., birth cohort) based on simulation studies. More details have been presented in **Supplementary Material S1**. As expected, selecting samples conditioning on genetic instruments G and outcome D of interest induces selection bias, with its impacts varying depending on the relative hazard of G and D on survival of the underlying population. Given summary statistics obtained from the original exposure and outcome GWASs, it seems not easy to recover the true causal estimate from the observed MR estimates in two-sample MR settings due to the essence of missing people before the recruitment of the original GWASs.

Validating MR Estimates by Detecting Selection Bias and Reproducing Associations of Interest Using Control Exposures

To explore selection bias, we reproduce a condition that does not involve the hypothesized causal mechanism but involves the same potential selection bias sources in the original MR study. We introduce an antagonist or decoy of E_1 as the control exposure E_2, mimicking a natural experiment, because E_2 acts as an endogenous intervention of E_1. Moreover, E_2 effects on survival would be nearly identical to E_1, as depicted in **Figures 1E,F**, but has an opposite impact on D from E_1. If such an E_2 exists, then any consistent effects of E_1 and E_2 on D would be mainly due to selection bias rather than study design. That is, the consistent effects of E_1 and E_2 on D could indicate potential selection bias. Otherwise, the estimated causal effects derived from the selected samples are robust to selection bias. Moreover, an intuitive interpretation herein is that the E_1–D association is credible and reproducible by using a relevant control exposure E_2 because of the known relationship between E_1 and E_2.

We can extend the selection of E_2 by using exposure with similar biological activity as E_1 because they are also likely to share the same potential selection bias sources and have similar or even the same effects on D. This idea is widely applied in developing pharmaceutical products [Food and Drug A (2014). Bioa, 2014; Committee for Medicinal P, 2010]. If such an E_2 exists, then any inconsistent effects of E_1 and E_2 on D would be mainly caused by potential selection bias. Conversely, consistent results of E_1 and E_2 on D would validate the estimated effects. In other words, these estimated effects derived from the selected samples are less likely to be affected by selection bias. Even if selection bias exists, its impact would be limited. It would not extend to reverse the causal direction or distort the estimated effect far away from the truth. Notably, the use of such kinds of control exposures does not require a null or

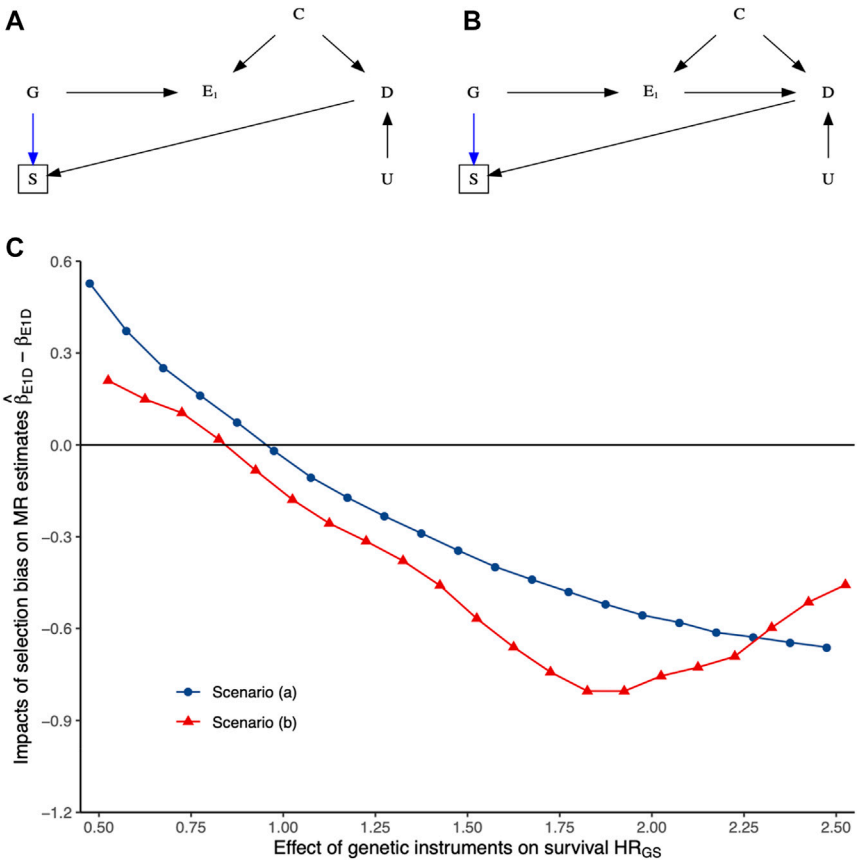

FIGURE 3 | The impacts of selection bias (i.e., $\hat{\beta}_{E1D} - \beta_{E1D}$) on two-sample Mendelian randomization (MR) estimates of the exposure E_1–outcome D association using the inverse-variance weighted method in terms of various relative hazards (HRs) of per-unit change in genetic instruments G (i.e., HR_{GS}) with a fixed effect of D (i.e., HR_{DS}) on survival of underlying population based on simulation studies, with more details presented in **Supplementary Material S1**. The upper panels **(A, B)** show scenarios that may happen in practice. The lower panel **(C)** shows the impacts of selection bias on MR estimates under each scenario. R codes for reproducing these results can be found in **Supplementary Material S2**.

well-established association between the control exposure E_2 and D.

Issue of Systematic Selection Bias

However, this method might still fail to detect selection bias if systematic selection bias exists, especially when E_1 and E_2 are selected from the same GWAS. In such a case, it might distort both the E_1–D and E_2–D associations similarly, such as reversing the estimated E_1–D and E_2–D associations simultaneously. To handle this situation, we introduced an additional negative (or positive) control E_3 with the same potential selection bias sources concerning the E_1–D association or identified a validated control exposure (E_2) that had a clear association with D to triangulate the estimated effects. As such, any associations of E_3/E_2 with D would indicate potentially systematic selection bias. Otherwise, the estimated effects derived from the selected samples are likely to be robust to selection bias and reproducible.

Choosing Control Exposures

Control exposures could be used to detect potential selection bias and validate MR estimates. To this end, it might be necessary to specify the criteria for choosing the control exposures E_2 and/or E_3 as follows.

1) The control exposure E_2 should have the same potential selection bias sources (e.g., affecting survival in the underlying population) as E_1 on D. For example, using antagonist, decoy, or an exposure with similar biological activity as E_2, such a criterion is approximately satisfied;
2) To explore potentially systematic selection bias, an additional control exposure (E_3) with the same potential selection bias sources as E_1 on D or a validated control exposure E_2 should have a well-established association with D.

We recommend choosing E_1, E_2, and/or E_3 from different GWASs to minimize potentially systematic selection bias. If such E_2 and E_3 exist, then the estimated effects of E_1, E_2, and E_3 on D can be used to detect potential selection bias and triangulate the causal estimates. The estimated E_1–D association would be more credible because it is robust to potential selection bias and can be reproducible using a relevant control exposure E_2 based on subject matter knowledge.

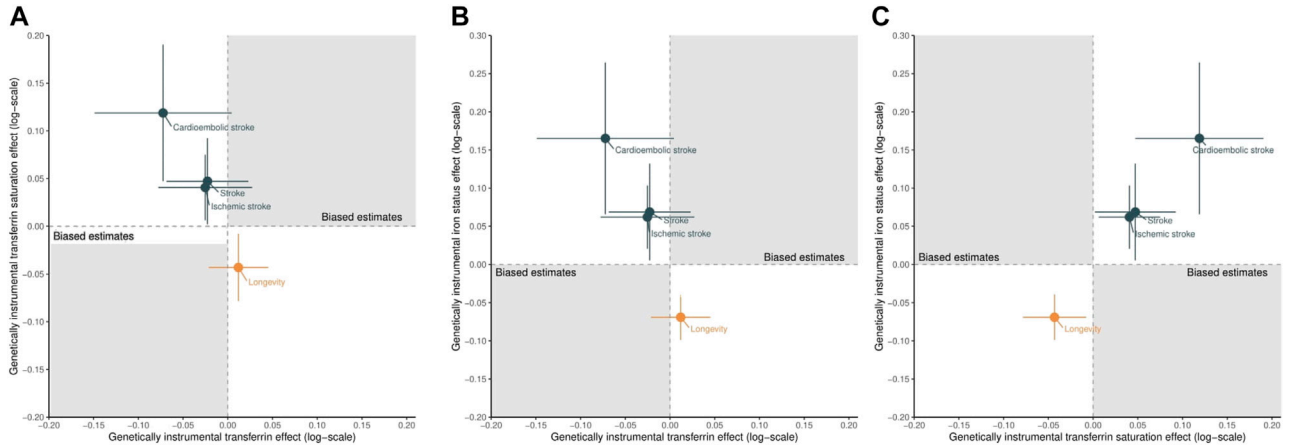

FIGURE 4 | Scatter plots of the estimated effects of genetically predicted higher transferrin versus higher transferrin saturation **(A)**, higher transferrin versus higher iron status **(B)**, and higher transferrin saturation versus higher iron status **(C)** on stroke (including ischemic and cardioembolic stroke) and longevity. Points located in the gray area indicate the presence of selection bias.

An Applied Example

To illustrate, we investigated the association of higher transferrin (i.e., E_1) with stroke (including ischemic and cardioembolic stroke), with transferrin saturation as a control exposure E_2 and iron status as a positive control exposure E_3. We selected transferrin saturation as the control exposure E_2 because it measures circulating iron and reflects the proportion of transferrin occupied by iron (Wish, 2006). Biologically, transferrin saturation is inversely associated with transferrin but positively associated with iron status. Furthermore, iron deficiency, reflected by lower transferrin saturation and higher transferrin, causes anemia and reduces lifespan directly or *via* competing risks [e.g., stroke (23), **Figure 2**] (McLean et al., 2009; Camaschella, 2015). Consequently, the associations of transferrin saturation and iron status with stroke are open to similar potential selection bias as the transferrin-stroke association. Hence, transferrin saturation and iron status are control exposures here. As such, any consistent transferrin–stroke and transferrin saturation–stroke associations (especially in the same causal direction) indicate potential selection bias. In addition, any null iron status–stroke association suggests the presence of systematic selection bias due to its clear associations with stroke and longevity (Gill et al., 2018; Daghlas and Gill, 2021); particularly, the iron status-longevity association is less likely to subject to selection bias (Andersen et al., 2012).

We selected independent ($r^2 < 0.01$) genetic instruments mimicking effects of transferrin (MR-base id: ieu-a-1052), transferrin saturation (MR-base id: ieu-a-1051), and iron status (MR-base id: ieu-a-1049) from the MR-base at a genome-wide significance $p < 5 \times 10^{-8}$ (Benyamin et al., 2014). We approximated the F statistics (i.e., the square of instrument's association on exposure divided by the square of its SE) to assess the instrument strength, where higher F statistics indicate a low risk of weak instrument bias (Bowden et al., 2016a). We excluded the instruments with F statistics less than 10 to alleviate potential weak instrument bias (Bowden et al., 2016a). We checked the shared instruments for transferrin, transferrin saturation, and iron status to explore the possibility of pleiotropic effects, but still used them in this example as they have been used similarly in a previous MR study (Daghlas and Gill, 2021). We further assessed associations of higher transferrin saturation and iron status with longevity, proxied by the heritable trait of parental lifespan from United Kingdom Biobank and LifeGen consortium (Timmers et al., 2019). Genetically predicted higher transferrin saturation and higher iron status were inversely associated with longevity, as shown in **Figure 4**, suggesting the similar or even the same selection bias sources as the transferrin–outcome association because it also appeared to affect longevity.

We applied the identified instruments to publicly available GWAS of European descent of stroke (40,585 cases and 406,111 controls), ischemic stroke (34,217 cases and 406,111 controls), and cardioembolic stroke (7,193 cases and 406,111 controls) (Timmers et al., 2019). **Supplementary Table S1** presents a detailed summary of the included studies. We extracted summary statistics for stroke (MR-base id: ebi-a-GCST005838), ischemic stroke (MR-base id: ebi-a-GCST005834), and cardioembolic stroke (MR-base id: ebi-a-GCST006910) from MR-base (Hemani et al., 2018). **Supplementary Table S2** lists genetic associations of the included instruments associated with stroke.

We assessed the associations of genetically predicted transferrin, transferrin saturation, and iron status with stroke using the Wald ratio (i.e., the ratio of the genetic outcome effect estimate and the corresponding genetic exposure effect estimate) or the inverse-variance weighted average of the Wald ratio estimates with random effects (Burgess et al., 2013). We assumed that all these associations were linear and homogeneous (Lawlor et al., 2008). We reported Cochran's Q-statistic to detect potential heterogeneity. We conducted sensitivity analyses using the weighted median (Bowden et al., 2016b), MR-Egger (Bowden et al., 2015), and MR-RAPS(40) to address the potential unknown pleiotropy statistically. We also

reported the MR-Egger intercept and its SE with *p*-value as an indicator of potential pleiotropy. Two-sided *p*-values at the Bonferroni-corrected threshold of 0.05/3 (for three exposures) = 0.017 were considered statistically significant. *P*-values between 0.017 and 0.05 were reported as nominal. Data involving these exemplars were publicly available, so it does not require ethical approval.

RESULTS

Up to 11 genetic instruments were used for transferrin (mean concentration 2.1 g/L and SD 0.43 g/L), 7 instruments for transferrin saturation (mean percentage 29.9% and SD 11.0%), and 5 instruments for iron status (mean concentration 18.4 μmol/L and SD 5.6 μmol/L). The F-statistics of instruments for transferrin ranged from 32.4 to 1,296.1, for transferrin saturation ranged from 35.6 to 808.5, and for iron status was 37.8 to 346.7, suggesting weak instrument bias to be less likely.

Figure 4 shows the scatter plot of the estimated effects of genetically predicted higher transferrin versus higher transferrin saturation (A), higher transferrin versus higher iron status (B), and higher transferrin saturation versus higher iron status (C) on stroke (including ischemic and cardioembolic stroke) and longevity, with full details presented in **Supplementary Table S3**. Genetically predicted higher transferrin was associated with a lower risk of stroke (**Figures 4A,B**), although these protective effects did not reach nominal significance ($p < 0.05$). Conversely, genetically predicted higher transferrin saturation was nominally associated with higher risk of stroke (**Figures 4A,C**). Such results suggest that the observed transferrin–stroke association is open to selection bias, possibly due to the missing people from the original GWAS of stroke because they died before recruitment from the genetic predictors of iron, an iron-related condition, stroke, or a competing risk of stroke, which attenuated the true association (**Figure 2**).

In addition, as expected (Gill et al., 2018; Daghlas and Gill, 2021), genetically predicted higher iron status was associated with increased stroke and reduced longevity, as shown in **Figures 4B,C** and **Supplementary Table S3**. Finally, the consistent effects of higher transferrin saturation and higher iron status on stroke and longevity further triangulated our conclusions. Even if selection bias exists, its impact on the transferrin saturation–stroke and iron status–stroke associations would be limited or at least could not reverse the observed associations or biased them to the null. These results support the advantages of using control exposures.

DISCUSSION

This paper advances the use of control exposures based on subject matter knowledge in MR studies to triangulate the estimated causal effects vulnerable to selection bias. The potential mechanisms underlying selection bias in MR lies in the re-opened backdoor pathway from genetic instruments to the outcome of interest in the selected samples. It violates the IV3 assumption and distorts the MR estimates. The applied example demonstrates that MR is vulnerable to selection bias because of missing data from sample selection (**Figures 1**, **3**), which is unlikely to be missing at random, so requires modeling of the missing data process to recover the estimates (Mohan and Pearl, 2021). Our proposal provides a valuable approach to assessing credible MR estimates in the presence of selection bias from selection of survivors.

Furthermore, the control exposures introduced in the proposal inherit properties similar to those of negative or positive control exposures used in the conventional observational studies but provide a more intuitive and clinically meaningful interpretation of the estimated effects (Lipsitch et al., 2010; Shi et al., 2020; Sanderson et al., 2021). Choosing antagonists, decoys, or exposures with similar biological activity as the control exposures based on subject matter knowledge may facilitate its application in MR studies. Systematic selection bias distorting both the exposure–outcome and control exposure–outcome associations, in a similar or even the same way, may exist, resulting in inconclusive or misleading conclusions. However, an additional or a validated control exposure with a clear association with the outcome provides another tool to triangulate the estimated effects. Notably, it is possible to use a single control exposure in the proposal solely to validate the MR estimates, especially when E_1, E_2, and E_3 are selected from different GWASs.

Despite the strengths of the proposal in validating MR estimates, limitations exist. First, the proposal only detects potential selection bias but fails to address it. The impact of selection bias on summary statistics obtained from the original GWAS might vary due to the small fraction of heritability explained by genetic variants and the small effect size of the genetic associations (Greenland, 2003; Freedman et al., 2004; Park et al., 2011; Schooling, 2019). Thus, the proposal might fail to detect its small effect on MR estimates. Nonetheless, routinely applying control exposures still adds more credibility to MR estimates. Second, the proposal inherits properties of the conventional MR; limitations such as the stringent instrumental assumptions remain (Davey Smith and Ebrahim, 2003; Smith and Ebrahim, 2004; Lawlor et al., 2008). However, recent advances in MR provide more tools to alleviate or even eliminate these limitations (Ye et al., 2019; Zhao et al., 2020; Liu et al., 2021). Third, choosing control exposures that have the same potential selection bias sources as the exposure–outcome association of interest or a clear association with the outcome might be difficult in practice, further limiting its application.

CONCLUSION

Routinely using control exposures in MR studies provides a helpful tool to validate estimated causal effects that are vulnerable to potential selection bias in the selected samples.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation was performed by ZY. The first draft of the article was written by ZY and all authors commented on previous versions of the article. All authors read and approved the final article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.729326/full#supplementary-material

REFERENCES

Andersen, P. K., Geskus, R. B., de Witte, T., and Putter, H. (2012). Competing Risks in Epidemiology: Possibilities and Pitfalls. *Int. J. Epidemiol.* 41 (3), 861–870. doi:10.1093/ije/dyr213

Arnold, B. F., Ercumen, A., Benjamin-Chung, J., and Colford, J. M., Jr (2016). Brief Report. *Epidemiology* 27 (5), 637–641. doi:10.1097/EDE.0000000000000504

Arnold, B. F., and Ercumen, A. (2016). Negative Control Outcomes. *JAMA* 316 (24), 2597–2598. doi:10.1001/jama.2016.17700

Bareinboim, E., and Pearl, J. (2012). *Controlling Selection Bias in Causal Inference*. Artificial Intelligence and Statistics (PMLR), 100–108.

Benyamin, B., Esko, T., Esko, T., Ried, J. S., Radhakrishnan, A., Vermeulen, S. H., et al. (2014). Novel Loci Affecting Iron Homeostasis and Their Effects in Individuals at Risk for Hemochromatosis. *Nat. Commun.* 5, 4926. doi:10.1038/ncomms5926

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016). Assessing the Suitability of Summary Data for Two-Sample Mendelian Randomization Analyses Using MR-Egger Regression: the Role of the I2 Statistic. *Int. J. Epidemiol.* 45 (6), 1961–1974. doi:10.1093/ije/dyw220

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44 (2), 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40 (4), 304–314. doi:10.1002/gepi.21965

Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. A., et al. (2020). Avoiding Dynastic, Assortative Mating, and Population Stratification Biases in Mendelian Randomization through Within-Family Analyses. *Nat. Commun.* 11 (1), 3519. doi:10.1038/s41467-020-17117-4

Burgess, S., Butterworth, A., Malarstig, A., and Thompson, S. G. (2012). Use of Mendelian Randomisation to Assess Potential Benefit of Clinical Intervention. *BMJ* 345, e7325. doi:10.1136/bmj.e7325

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian Randomization Analysis with Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* 37 (7), 658–665. doi:10.1002/gepi.21758

Burgess, S., Scott, R. A., Scott, R. A., Timpson, N. J., Davey Smith, G., and Thompson, S. G. (2015). Using Published Data in Mendelian Randomization: a Blueprint for Efficient Identification of Causal Risk Factors. *Eur. J. Epidemiol.* 30 (7), 543–552. doi:10.1007/s10654-015-0011-z

Camaschella, C. (2015). Iron-deficiency Anemia. *N. Engl. J. Med.* 372 (19), 1832–1843. doi:10.1056/NEJMra1401038

Canan, C., Lesko, C., and Lau, B. (2017). Instrumental Variable Analyses and Selection Bias. *Epidemiology* 28 (3), 396–398. doi:10.1097/EDE.0000000000000639

Chang, Y.-L., Hung, S.-H., Ling, W., Lin, H.-C., Li, H.-C., and Chung, S.-D. (2013). Association between Ischemic Stroke and Iron-Deficiency Anemia: a Population-Based Study. *PLoS One* 8 (12), e82952. doi:10.1371/journal.pone.0082952

Committee for Medicinal Products for Human U (2010). *Guideline on the Investigation of Bioequivalence*. London: European Medicines Agency, 1–27.

Daghlas, I., and Gill, D. (2021). Genetically Predicted Iron Status and Life Expectancy. *Clin. Nutr.* 40 (4), 2456–2459. doi:10.1016/j.clnu.2020.06.025

Davey Smith, G., and Ebrahim, S. (2003). 'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease?*. *Int. J. Epidemiol.* 32 (1), 1–22. doi:10.1093/ije/dyg070

Davies, N. M., Holmes, M. V., and Davey Smith, G. (2018). Reading Mendelian Randomisation Studies: a Guide, Glossary, and Checklist for cliniciansPMCPMC6041728 Interests and Declare that We Have No Competing Interests. *BMJ* 362, k601. doi:10.1136/bmj.k601

Diseases, G. B. D., and Injuries, C. (2020). Global burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990-2019: a Systematic Analysis for the Global Burden of Disease Study 2019. *Lancet* 396 (10258), 1204–1222. doi:10.1016/S0140-6736(20)30925-9

Fishbane, S., Pollack, S., Feldman, H. I., and Joffe, M. M. (2009). Iron Indices in Chronic Kidney Disease in the National Health and Nutritional Examination Survey 1988-2004. *Cjasn* 4 (1), 57–61. doi:10.2215/CJN.01670408

Food, Drug A (2014). *Bioavailability and Bioequivalence Studies Submitted in NDAs or INDs—General Considerations (Draft Guidance)*. Silver Spring (MD): Food and Drug Administration.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., et al. (2004). Assessing the Impact of Population Stratification on Genetic Association Studies. *Nat. Genet.* 36 (4), 388–393. doi:10.1038/ng1333

Gill, D., Benyamin, B., Moore, L. S. P., Monori, G., Zhou, A., Koskeridis, F., et al. (2019). Associations of Genetically Determined Iron Status across the Phenome: A Mendelian Randomization Study. *Plos Med.* 16 (6), e1002833, 2019 . PMCPMC6586257 following competing interests: LSPM has consulted for bioMerieux (2014), DNAelectronics (2015-2018), Dairy Crest (2017-2018), and Pfizer (2018) and has received research grants from Leo Pharma (2016) and educational support from Eumedica (2016-2017). All other authors have no competing interest to declare. doi:10.1371/journal.pmed.1002833

Gill, D., Del Greco M., F., Walker, A. P., Srai, S. K. S., Laffan, M. A., and Minelli, C. (2017). The Effect of Iron Status on Risk of Coronary Artery Disease. *Arterioscler Thromb. Vasc. Biol.* 37 (9), 1788–1792. doi:10.1161/ATVBAHA.117.309757

Gill, D., Monori, G., Tzoulaki, I., and Dehghan, A. (2018). Iron Status and Risk of Stroke. *Stroke* 49 (12), 2815–2821. doi:10.1161/STROKEAHA.118.022701

Gillum, R. F., Sempos, C. T., Makuc, D. M., Looker, A. C., Chien, C.-Y., and Ingram, D. D. (1996). Serum Transferrin Saturation, Stroke Incidence, and Mortality in Women and Men: The NHANES I Epidemiologic Followup Study. *Am. J. Epidemiol.* 144 (1), 59–68. doi:10.1093/oxfordjournals.aje.a008855

Gkatzionis, A., and Burgess, S. (2019). Contextualizing Selection Bias in Mendelian Randomization: How Bad Is it Likely to Be. *Int. J. Epidemiol.* 48 (3), 691–701. doi:10.1093/ije/dyy202

Global Burden of Disease Collaborative Network (2020). *Global Burden of Disease Study 2019 (GBD 2019) Results*. Seattle, United States: Institute for Health Metrics and Evaluation IHME. Available at: http://ghdx.healthdata.org/gbd-results-tool (Accessed August 18, 2021).

Greenland, S. (2003). Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias. *Epidemiology* 14 (3), 300–306. doi:10.1097/01.ede.0000042804.12056.6c

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base Platform Supports Systematic Causal Inference across the Human Phenome. *Elife* 7, 7. doi:10.7554/eLife.34408

Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the Causal Relationship between Imprecisely Measured Traits Using GWAS Summary Data. *Plos Genet.* 13 (11), e1007081. doi:10.1371/journal.pgen.1007081

Kesteloot, H., and Decramer, M. (2008). Age at Death from Different Diseases: the Flemish Experience during the Period 2000-2004. *Acta Clinica Belgica* 63 (4), 256–261. doi:10.1179/acb.2008.047

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology. *Statist. Med.* 27 (8), 1133–1163. doi:10.1002/sim10.1002/sim.3034

Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. (2010). Negative Controls. *Epidemiology* 21 (3), 383–388. doi:10.1097/EDE.0b013e3181d61eeb

Liu, Z., Ye, T., Sun, B., Schooling, M., and Tchetgen, E. T. (2021). On Mendelian Randomisation Mixed-Scale Treatment Effect Robust Identification (MR MiSTERI) and Estimation for Causal Inference. *medRxiv* 2020, 20204420. doi:10.1101/2020.09.29.20204420

Marniemi, J., Alanen, E., Impivaara, O., Seppänen, R., Hakala, P., Rajala, T., et al. (2005). Dietary and Serum Vitamins and Minerals as Predictors of Myocardial Infarction and Stroke in Elderly Subjects. *Nutr. Metab. Cardiovasc. Dis.* 15 (3), 188–197. doi:10.1016/j.numecd.2005.01.001

McLean, E., Cogswell, M., Egli, I., Wojdyla, D., and de Benoist, B. (2009). Worldwide Prevalence of Anaemia, WHO Vitamin and Mineral Nutrition Information System, 1993-2005. *Public Health Nutr.* 12 (4), 444–454. doi:10.1017/S1368980008002401

Menotti, A., Puddu, P. E., Tolonen, H., Adachi, H., Kafatos, A., and Kromhout, D. (2019). Age at Death of Major Cardiovascular Diseases in 13 Cohorts. The Seven Countries Study of Cardiovascular Diseases 45-year Follow-Up. *Acta Cardiologica* 74 (1), 66–72. doi:10.1080/00015385.2018.1453960

Mohan, K., and Pearl, J. (2021). Graphical Models for Processing Missing Data. *J. Am. Stat. Assoc.* 116 (534), 1023–1037. doi:10.1080/01621459.2021.1874961

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., and Davey Smith, G. (2018). Collider Scope: when Selection Bias Can Substantially Influence Observed Associations. *Int. J. Epidemiol.* 47 (1), 226–235. doi:10.1093/ije/dyx206

Munafò, M., and Smith, G. D. (2018). Biased Estimates in Mendelian Randomization Studies Conducted in Unrepresentative Samples. *JAMA Cardiol.* 3 (2), 181. doi:10.1001/jamacardio.2017.4279

Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B. L., Whittaker, J. C., and Leon, D. A. (2006). Limits to Causal Inference Based on Mendelian Randomization: A Comparison with Randomized Controlled Trials. *Am. J. Epidemiol.* 163 (5), 397–403. doi:10.1093/aje/kwj062

Park, J.-H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., et al. (2011). Distribution of Allele Frequencies and Effect Sizes and Their Interrelationships for Common Genetic Susceptibility Variants. *Proc. Natl. Acad. Sci.* 108 (44), 18026–18031. doi:10.1073/pnas.1114759108

Sanderson, E., Richardson, T. G., Hemani, G., and Davey Smith, G. (2021). The Use of Negative Control Outcomes in Mendelian Randomization to Detect Potential Population Stratification. *Int. J. Epidemiol.* doi:10.1093/ije/dyaa288

Schooling, C. M. (2019). *Biases in GWAS – the Dog that Did Not Bark*. bioRxiv, 709063. doi:10.1101/709063

Schooling, C. M., Lopez, P. M., Yang, Z., Zhao, J. V., Au Yeung, S. L., and Huang, J. V. (2020). Use of Multivariable Mendelian Randomization to Address Biases Due to Competing Risk before Recruitment. *Front. Genet.* 11, 610852. doi:10.3389/fgene.2020.610852

Sekula, P., Del Greco M, F., Pattaro, C., and Köttgen, A. (2016). Mendelian Randomization as an Approach to Assess Causality Using Observational Data. *Jasn* 27 (11), 3253–3265. doi:10.1681/ASN10.1681/asn.2016010098

Shan, Y., Lambrecht, R. W., and Bonkovsky, H. L. (2005). Association of Hepatitis C Virus Infection with Serum Iron Status: Analysis of Data from the Third National Health and Nutrition Examination Survey. *Clin. Infect. Dis.* 40 (6), 834–841. doi:10.1086/428062

Shi, X., Miao, W., and Tchetgen, E. T. (2020). A Selective Review of Negative Control Methods in Epidemiology. *Curr. Epidemiol. Rep.*, 1–13. doi:10.1007/s40471-020-00243-4

Smit, R. A. J., Trompet, S., Dekkers, O. M., Jukema, J. W., and le Cessie, S. (2019). Survival Bias in Mendelian Randomization Studies. *Epidemiology* 30 (6), 813–816. doi:10.1097/EDE.0000000000001072

Smith, G. D., and Ebrahim, S. (2004). Mendelian Randomization: Prospects, Potentials, and Limitations. *Int. J. Epidemiol.* 33 (1), 30–42. doi:10.1093/ije/dyh132

Staiger, D., and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica* 65 (3), 557–586. doi:10.2307/2171753

Tchetgen Tchetgen, E. J., and Wirth, K. E. (2017). A General Instrumental Variable Framework for Regression Analysis with Outcome Missing Not at Random. *Biom* 73 (4), 1123–1131. doi:10.1111/biom.12670

Timmers, P. R., Mounier, N., Lall, K., Fischer, K., Ning, Z., Feng, X., et al. (2019). Genomics of 1 Million Parent Lifespans Implicates Novel Pathways and Common Diseases and Distinguishes Survival Chances. *Elife* 8, 8. doi:10.7554/eLife.39856

van der, D. L., Grobbee, D. E., Roest, M., Marx, J. J. M., Voorbij, H. A., and van der Schouw, Y. T. (2005). Serum Ferritin Is a Risk Factor for Stroke in Postmenopausal Women. *Stroke* 36 (8), 1637–1641. doi:10.1161/01.STR.0000173172.82880.72

VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M., and Kraft, P. (2014). Methodological Challenges in Mendelian Randomization. *Epidemiology* 25 (3), 427–435. doi:10.1097/EDE.0000000000000081

Vansteelandt, S., Dukes, O., and Martinussen, T. (2018). Survivor Bias in Mendelian Randomization Analysis. *Biostatistics* 19 (4), 426–443. doi:10.1093/biostatistics/kxx050

Vansteelandt, S., Walter, S., and Tchetgen Tchetgen, E. (2018). Eliminating Survivor Bias in Two-Stage Instrumental Variable Estimators. *Epidemiology* 29 (4), 536–541. doi:10.1097/EDE.0000000000000835

Wang, K., and Han, S. (2021). Effect of Selection Bias on Two Sample Summary Data Based Mendelian Randomization. *Sci. Rep.* 11 (1), 7585. doi:10.1038/s41598-021-87219-6

Wish, J. B. (2006). Assessing Iron Status: beyond Serum Ferritin and Transferrin Saturation. *Cjasn* 1 (1 Suppl. l), S4–S8. doi:10.2215/CJN.01490506

Yang, Z., Schooling, C. M., and Kwok, M. K. (2021). Genetic Evidence on the Association of Interleukin (IL)-1-mediated Chronic Inflammation with Airflow Obstruction: A Mendelian Randomization Study. *COPD: J. Chronic Obstructive Pulm. Dis.* 18 (4), 432–442. doi:10.1080/15412555.2021.1955848

Ye, T., Shao, J., and Kang, H. (2019). Debiased Inverse-Variance Weighted Estimator in Two-Sample Summary-Data Mendelian Randomization. *Ann. Stat.* 49 (4), 2079–2100. doi:10.1214/20-AOS2027

Yuan, S., and Larsson, S. (2020). Causal Associations of Iron Status with Gout and Rheumatoid Arthritis, but Not with Inflammatory Bowel Disease. *Clin. Nutr.* 39 (10), 3119–3124. doi:10.1016/j.clnu.2020.01.019

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020). Statistical Inference in Two-Sample Summary-Data Mendelian Randomization Using Robust Adjusted Profile Score. *Ann. Statist* 48 (3), 1742–1769. doi:10.1214/19-AOS1866

Zhao, Q., Chen, Y., Wang, J., and Small, D. S. (2019). Powerful Three-Sample Genome-wide Design and Robust Statistical Inference in Summary-Data Mendelian Randomization. *Int. J. Epidemiol.* 48 (5), 1478–1492. doi:10.1093/ije/dyz142

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 November 2021
doi: 10.3389/fgene.2021.724785

Exploration of Potential miRNA Biomarkers and Prediction for Ovarian Cancer Using Artificial Intelligence

Farzaneh Hamidi[1], Neda Gilani[1], Reza Arabi Belaghi[2,3], Parvin Sarbakhsh[1], Tuba Edgünlü[4] and Pasqualina Santaguida[5]*

[1]Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran, [2]Department of Statistics, Faculty of Mathematical Science, University of Tabriz, Tabriz, Iran, [3]Department of Mathematics, Applied Mathematics and Statistics, Uppsala University, Uppsala, Sweden, [4]Department of Medical Biology, Faculty of Medicine, Muğla Sıtkı Koçman University, Muğla, Turkey, [5]Department of Health Research and Methods, McMaster University, Hamilton, ON, Canada

Ovarian cancer is the second most dangerous gynecologic cancer with a high mortality rate. The classification of gene expression data from high-dimensional and small-sample gene expression data is a challenging task. The discovery of miRNAs, a small non-coding RNA with 18–25 nucleotides in length that regulates gene expression, has revealed the existence of a new array for regulation of genes and has been reported as playing a serious role in cancer. By using LASSO and Elastic Net as embedded algorithms of feature selection techniques, the present study identified 10 miRNAs that were regulated in ovarian serum cancer samples compared to non-cancer samples in public available dataset GSE106817: hsa-miR-5100, hsa-miR-6800-5p, hsa-miR-1233-5p, hsa-miR-4532, hsa-miR-4783-3p, hsa-miR-4787-3p, hsa-miR-1228-5p, hsa-miR-1290, hsa-miR-3184-5p, and hsa-miR-320b. Further, we implemented state-of-the-art machine learning classifiers, such as logistic regression, random forest, artificial neural network, XGBoost, and decision trees to build clinical prediction models. Next, the diagnostic performance of these models with identified miRNAs was evaluated in the internal (GSE106817) and external validation dataset (GSE113486) by ROC analysis. The results showed that first four prediction models consistently yielded an AUC of 100%. Our findings provide significant evidence that the serum miRNA profile represents a promising diagnostic biomarker for ovarian cancer.

Keywords: Biomarker, Elasticnet, Feature Selection, Gene Expression Omnibus (GEO), Lasso, Machine Learning, Ovarian Cancer

OPEN ACCESS

Edited by:
Tao Wang,
Northwestern Polytechnical University, China

Reviewed by:
Bor-Sen Chen,
National Tsing Hua University, Taiwan
Yinghui Zhao,
Second Hospital of Shandong University, China

****Correspondence:***
Reza Arabi Belaghi
r.arabi@tabrizu.ac.ir

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *14 June 2021*
Accepted: *07 October 2021*
Published: *25 November 2021*

Citation:
Hamidi F, Gilani N, Belaghi RA, Sarbakhsh P, Edgünlü T and Santaguida P (2021) Exploration of Potential miRNA Biomarkers and Prediction for Ovarian Cancer Using Artificial Intelligence. Front. Genet. 12:724785. doi: 10.3389/fgene.2021.724785

INTRODUCTION

Ovarian cancer is a major clinical challenge in gynecologic oncology. Due to the lack of a proper biomarker-based screening method, most patients are asymptomatic until the disease has metastasized and two-thirds of patients are diagnosed with advanced stages (Lheureux et al., 2019). The International Federation of Gynecology and Obstetrics (FIGO) reported that in the majority of those diagnosed in stage three or four ovarian cancer (2014), more than 70% will have a relapse of their disease within the first 5 years (Reid et al., 2017). Currently, there is an acute need to know potential biomarkers that could lead to the growth of modern and more accurate predictors for ovarian cancer diagnosis and prognosis. As noted, one of the most common gynecologic malignancy is epithelial ovarian cancer (EOC), with each year of about 230,000 new cases and almost 140,000

deaths (Greenlee et al., 2001). In 2020, it is estimated that approximately 21,750 new cases and 13,940 deaths occurred in the United States and 29,000 deaths happened in Europe due to ovarian cancer (Iorio et al., 2007). Therefore, the underlying molecular mechanism has not yet been elucidated. The timely prediction of ovarian cancer would benefit women, healthcare systems, and society as a whole. Accurate and reliable prediction models would enable preventative interventions to reduce the morbidity and mortality associated with ovarian cancer (Harter et al., 2008).

MicroRNAs

MicroRNAs (miRNA) are important genomic datasets in the human genome that play a regulative impress in cellular processes. miRNAs are a type of non-coding RNA with 18–25 nucleotides in length and reported to play a serious role in human cancers. miRNAs are often copied from DNA sequences to primary miRNAs. Subsequent processes lead to the production of precursor miRNAs and mature miRNAs. The most common mode of action of miRNAs is their interaction with the 3′ untranslated region (3′ UTR) of target mRNAs and increased mRNA degradation and translation suppression. miRNAs can also interact with the five UTR, coding sequence, and promoter regions of their target. In some cases, miRNA interaction with target sequences can induce transcription or regulate transcription. Various parameters modulate miRNA-mRNA interaction, including the subcellular state of miRNAs, the amount of miRNAs and target mRNAs, and the affinity of the interactions (Chen et al., 2015). miRNAs play a role in almost all aspects of cancer biology, such as apoptosis, proliferation, metastasis, and angiogenesis (Lee and Dutta, 2009). In addition, miRNAs have been proposed as potential biomarkers for the recognition of various different cancer types (Lin et al., 2015). Some studies also reported that several miRNAs have a potential value as diagnostic biomarkers of ovarian cancer (Banka and Dara, 2012; Yao et al., 2020).

Related Works

The down-regulation of miRNAs was found to be related to the progression and the prognoses of cancers. Falzone et al. determined that a group of 16 miRNAs were significantly expressed between bladder cancer patients and normal samples; they serve to modulate the expression of both EMT and NGAL/MMP-9 pathways (Falzone et al., 2016). Falzone et al.

TABLE 1 | Summary of miRNA genes shown to be statistically significantly associated with ovarian cancer.

Reference	Association	Up-regulated miRNA	Down-regulated miRNA
Tuncer et al. (2020)	Epithelial ovarian cancer	miR-6131, miR-1305, miR-197-3p, and miR-3651	miR-3135b, miR-4430, miR-664b-5p, and miR-766-3p
Nam et al. (2008)	Serous ovarian cancer	miR-16, miR-20a, miR-21, and miR-27a	miR-145, miR-125B, miR-125B, and miR-100
Iorio et al. (2007)	Epithelial ovarian cancer and normal	miR-200a, miR-141, miR-200c, miR-200b, miR-182, and miR-205	miR-127, miR-140, miR-9, miR-101, miR-147, miR-204, miR-211, miR-124a, and miR-302b

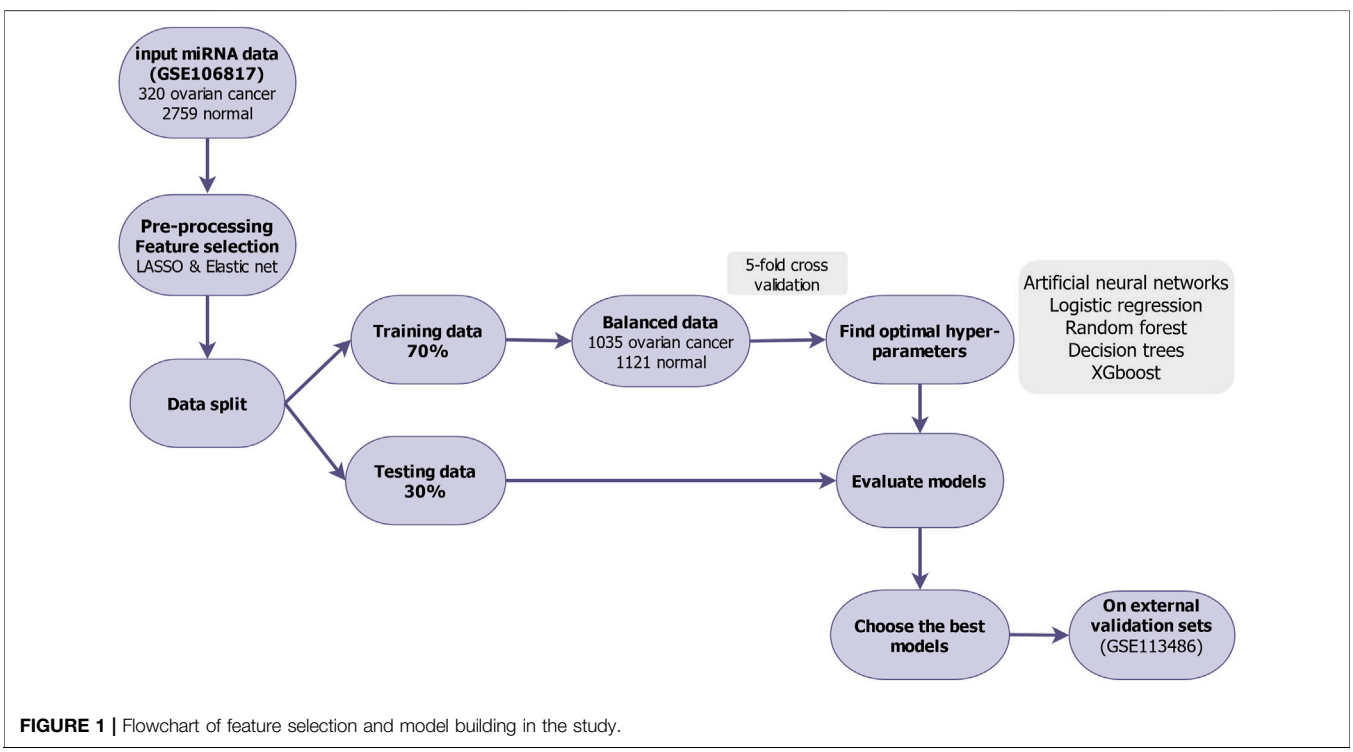

FIGURE 1 | Flowchart of feature selection and model building in the study.

TABLE 2 | miRNAs identified with threshold over 80% importance in both Lasso and Elastic net in the dataset GSE106817 with miRNA status.

miRNA-ID List	Importnace in Elastic Net	Importnace in LASSO (%)	adj.*p*-value	B	logFC	miRNAStatus
hsa-miR-5100	100	100	<0.001	16.18	4.15	Upregulated
hsa-miR-1290	100	100	<0.001	13.00	5.61	Upregulated
hsa-miR-320b	—	88.07	<0.001	12.25	4.11	Upregulated
hsa-miR-1233-5p	85.63	87.81	<0.001	11.78	2.36	Upregulated
hsa-miR-4783-3p	100	87.44	<0.001	10.36	2.89	Upregulated
hsa-miR-6800-5p	—	84.07	<0.001	8.66	−1.60	Downregulated
hsa-miR-4532	85.51	—	<0.001	6.95	2.90	Upregulated
hsa-miR-3184-5p	83.33	—	<0.001	5.29	−3.23	Downregulated
hsa-miR-4787-3p	100	—	<0.001	3.82	2.30	Upregulated
hsa-miR-1228-5p	88.83	—	<0.001	2.03	−0.93	Downregulated

TABLE 3 | Predictive power of models for ovarian cancer classification and prediction in the external (GSE113486) validation data.

Classifier	Hyperparameters	AUC[a] (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Negative predictive value (%)	Positive predictive value (%)	Kappa (%)
LR	Parameters[b]	100	100	100	100	100	100	100
DT	Cp[c] = 0.0115942	92.60	91.30	92.50	90.38	88.10	94	82.41
RF	Mtry[d] = 2	100	97.83	95	100	100	96.30	95.55
ANN	Size[e] = 3 and decay[f] = 1e–04	100	100	100	100	100	100	100
XGB	nrounds = 50, max_depth[g] = 2, eta = 0.3, gamma[h] = 0, colsample_bytree[i] = 0.8, min_child_weight[j] = 1 and subsample[k] = 1	100	98.91	97.50	100	100	98.11	97.78

[a]*The area under the receiver operating characteristic curve (maximum) was used to select the optimal model.*

[b]*The formula for logistic regression for prediction of ovarian cancer is* $p = (1 + e^{-[14.19-40.34(has.miR.6800.5p)+3.61(has.miR.1228.5p)+16.09(has.miR.5100)+2.86(has.miR.1290)+4.17(has.miR.4783.3p)-8.9(has.miR.3184.5p)+8(has.miR.320b)+9.23(has.miR.4532)-4.2(has.miR.4787.3p)-0.65(has.miR.1233.5p)]})^{-1}$.

[c]*The complexity parameter (cp) is used to control the size of the decision tree and to select the optimal tree size. If the cost of adding an additional variable to the decision tree from the current node is above the value of the cp, then tree building does not continue.*

[d]*mtry is the number of variables available for splitting at each tree node. In the random forests literature, this is referred to as the mtry parameter.*

[e]*Size is the number of units in a hidden layer.*

[f]*Decay is the regularization parameter used to avoid over-fitting.*

[g]*max-depth is used to control over-fitting as higher depth will allow model to learn relations very specific to a particular sample.*

[h]*gamma A node is split only when the resulting split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split. Makes the algorithm conservative. The values can vary depending on the loss function and should be tuned.*

[i]*Denotes the fraction of columns to be randomly sampled for each tree.*

[j]*min_child_weight used to control over-fitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample selected for a tree. Too high values can lead to under-fitting; hence, it should be tuned using CV.*

[k]*Subsample lower values make the algorithm more conservative and prevent overfitting, but too small values might lead to under-fitting.*

identified a series of novel microRNAs and their diagnostic and prognostic significance in oral cancer and their study has therefore developed a molecular detector (Falzone et al., 2019). Another study by Asano et al. reported circulating serum miRNA profile classifier for the detection of sarcoma samples using seven miRNAs (Asano et al., 2019). **Table 1** summarizes the results of miRNA associations with ovarian cancer in three recent genetic biomarker studies.

MATERIALS AND METHODS

Candidate Genetic Biomarkers

To identify a robust circulating miRNA biomarker, we searched the Gene Expression Omnibus (GEO) database with specific keywords, namely, ["ovarian neoplasms" (MeSH Terms) OR ovarian cancer (All Fields)] AND "*Homo sapiens*" (porgn) AND ["microRNAs" (MeSH Terms) OR miRNA (All Fields)]. Then, two datasets using the same platform (3D-Gene Human miRNA V21_1.0.0) with larger sample sizes GSE106817 and GSE113486 were included (360 ovarian cancer patients and 2,811 non-cancer controls in total) for our analysis. GSE106817 (320 ovarian cancer patients and 2,759 non-cancer controls) was used as the internal discovery cohort, and GSE113486 (40 ovarian cancer patients and 52 non-cancer controls) was used for independent validation. This study was approved by the Ethics Committee of Tabriz University of Medical Sciences (No: IR. TBZMED.REC.1400.006).

Data Preprocessing

Our analytical process is summarized in **Figure 1**. To discover biomarkers for ovarian cancer, the free available dataset GSE106817 includes 320 ovarian cancer patients and 2,759 non-cancer controls (11% ovarian cancer and 89% non-cancer). For machine learning analysis purpose, we preprocessed, cleaned, and then normalized by min-max normalization the data (Huang J. et al., 2015).

Feature Selection Algorithms

Feature (variable) selection is the main phase for selecting biomarkers in biological data with high dimension and small sample (p > n). Regularization is a kind of various technique of feature selection methods that use different penalty function to reduce the risk of overfitting and also reduce the complexity of the models (Drotár et al., 2015). Least Absolute Shrinkage and Selection Operation (LASSO) and Elastic Net are the most common embedded feature selection method which are an alternative to the subset selection and dimension reduction techniques. Thus, these algorithms can significantly reduce the variance by performing the variable selection. In the first phase, the expression levels of all 2,568 miRNAs from GSE106817 were analyzed to identify miRNAs as the candidate biomarkers by LASSO and Elastic Net (Zou and Hastie, 2005). For this sake, we used the "glmnet" package in R version 4.0.3. The next subsection gives a brief introduction to the LASSO and Elastic-Net.

LASSO

LASSO has been proposed by Tibshirani (Hastie et al., 2009) for parameter estimation and variable selection simultaneously in regression analysis. LASSO is a special instance of the penalized least squares regression with L1-penalty function. LASSO estimate of β can be defined as

$$\hat{\beta}_{la}(\lambda) = argmin_{\beta}\left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda\|\beta\|_1\right);$$

Where

$$\|Y - X\beta\|_2^2 = \sum_{i=0}^{n}(Y_i - \beta_i X_i)^2,\ \|\beta\|_1 = \sum_{j=1}^{k}|\beta_j| \text{ and } \lambda \geq 0.$$

Elastic Net

Elastic Net (ENET) is a convex combination of Ridge and LASSO which shrinks some coefficients to be very small, and on the other hand, similar to the LASSO, ENET set some coefficients to be exactly zero. Elastic Net is an extension of the LASSO that is robust to extreme correlations among the predictors (Zou and Hastie, 2005). When the number of variables exceeds the number of instances (p > n), ENET performs better than LASSO. To trim the instability of the LASSO solution paths, when predictors are highly correlated, the Elastic Net was proposed for analyzing high dimensional data (Liang and Jacobucci, 2020). The Elastic Net uses a mixture of the LASSO and ridge regression penalties and can be formulated as:

$$\hat{\beta}_{el}(\lambda) = \text{argmin}_{\beta}\left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1\right)$$
$$\text{and } \lambda_1, \lambda_2 \geq 0,\ \lambda_1 + \lambda_2 = 1.$$

The entire path of variable selection by LASSO and ENET algorithms is computed by the path coordinate descent algorithms which is available "glmnet" package in R (Friedman et al., 2010).

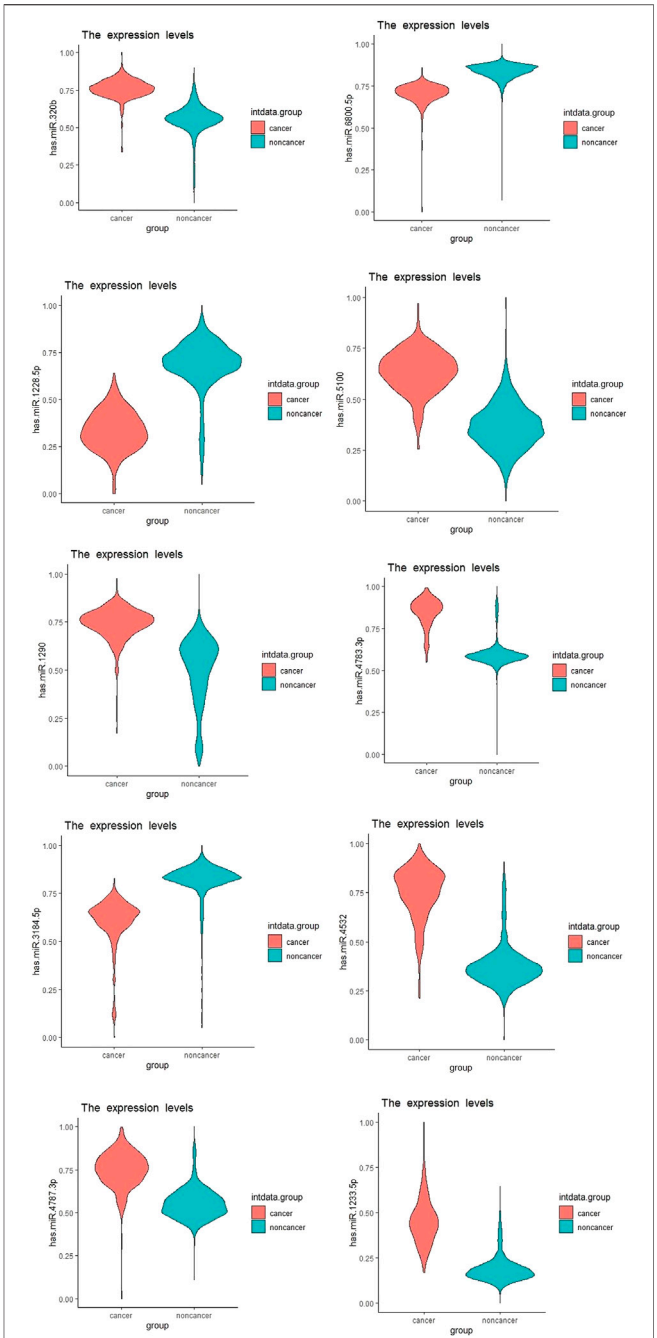

FIGURE 2 | Boxplots of the 10 identified miRNAs in ovarian cancer patients compared with the non-cancer control patients in the dataset GSE106817.

Machine Learning Classifier

Over the last decade, machine learning has been used for successful classification, both for identifying specific classes and for diagnosing cancers (Wang et al., 2005). We use this approach to characterize miRNAs with biomarker potential that could be useful for the diagnosis and/or prognosis of ovarian cancer for potential benefit for public health (screening) and for reduction in economic burden (Deb et al., 2018).

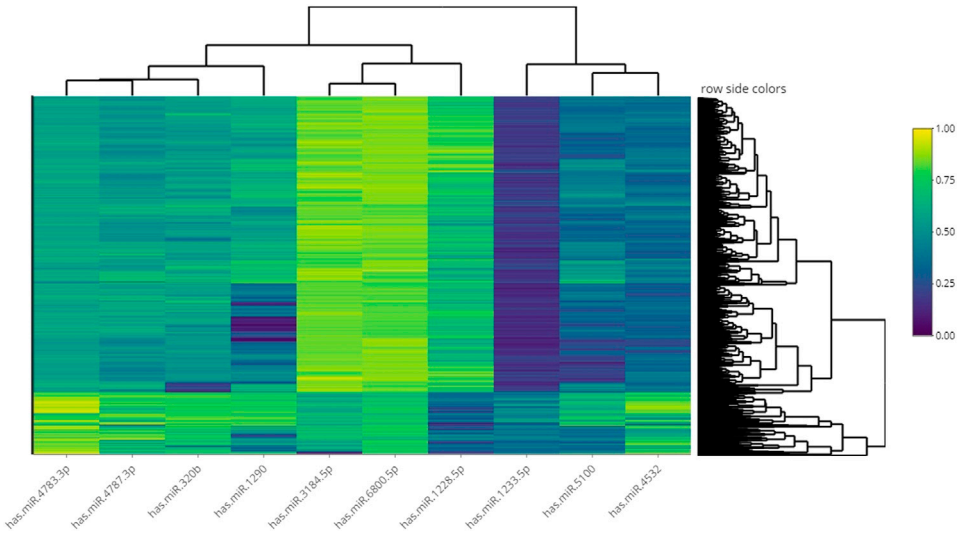

FIGURE 3 | Heatmap of hierarchical clustering analysis using the 10 identified miRNAs to distinguish different samples in the dataset GSE106817.

Logistic Regression

Logistic regression (LR) analyzes the relationship among multiple independent variables and a univariate binary outcome variable (Menard, 2010). One of the main advantage of the logistic regression is its simplicity and interpretability by providing the odds ratio for an outcome (Stoltzfus, 2011). The goodness of fit of a logistic regression model is evaluated using the area under the curve (AUC) (Abdulqader, 2017).

Artificial Neural Networks

Artificial neural networks (ANN) have been broadly used in medical studies (DeGregory et al., 2018). Such algorithms perform well when there are complex and non-linear associations between variables (Hassanipour et al., 2019). Briefly, artificial neural networks use predictors as inputs and connect them to multiple hidden layer combinations by assigning suitable weights to predict the outcome (Lisboa and Taktak, 2006). The hidden layers and weights must be appropriately selected by the analyst (Sherriff et al., 2004).

Decision Trees

Decision trees (DT) (Hassanipour et al., 2019) are a type of supervised machine learning that can be used to find attributes and extract patterns from big databases that are important for predictive modeling (Lisboa and Taktak, 2006). Decision trees are the most direct forward algorithm that processes a visual representation of the relationships between the independents and dependent variables (Hassanipour et al., 2019). However, the variation in the decision trees, in some instances, can be improved by using random forests for the outcomes of randomly generated decision trees to produce a more robust model (Vens et al., 2008).

Random Forest

Among several machine learning algorithms, random forest (RF) has a number of interesting characteristics. Firstly, RF does not overfit when the number of features exceeds the number of instances. Secondly, it does feature selection implicitly. Thirdly, it takes into account the interactions between variables (Okun and Priisalu, 2007). RF is an instance of ensemble learning, in which a complex model is made by combining many simple decision tree models to decrease the variance (Qi, 2012).

XGBoosting

XGBoost (XGB) abbreviated for extreme Gradient Boosting package. XGB is a decision-tree-based ensemble of machine learning algorithms that uses a scalable implementation of gradient boosting XGB framework tree boosting (Chen et al., 2015). The most significant component in XGB success is its scalability across all scenarios which is due to a number of major systems and algorithmic enhancements (Chen and Guestrin, 2016).

Training Machine Learning Models and Hyper Parameter Setting

We started by removing the noise variables with LASSO and ENET. We then implemented SMOTE random oversampling techniques to balance cancer and non-cancer cases in the training data (GSE106817) using the "ROSE" package (Lunardon et al., 2014). We find the optimal prediction models in the training data by using 5-fold cross-validation. We performed ovarian cancer classification using ANN, LR, RF, DT, and XGB (James et al., 2013) algorithms to build our models, after finalizing the optimal hyperparameters for each model. The varImp () function in the *caret* package was used to determine the miRNAs that are the most important. In this, study we select the most important variables (variable importance >80%) from each of the models. We evaluated our model prediction performances based on several measures of accuracy, including sensitivity, specificity,

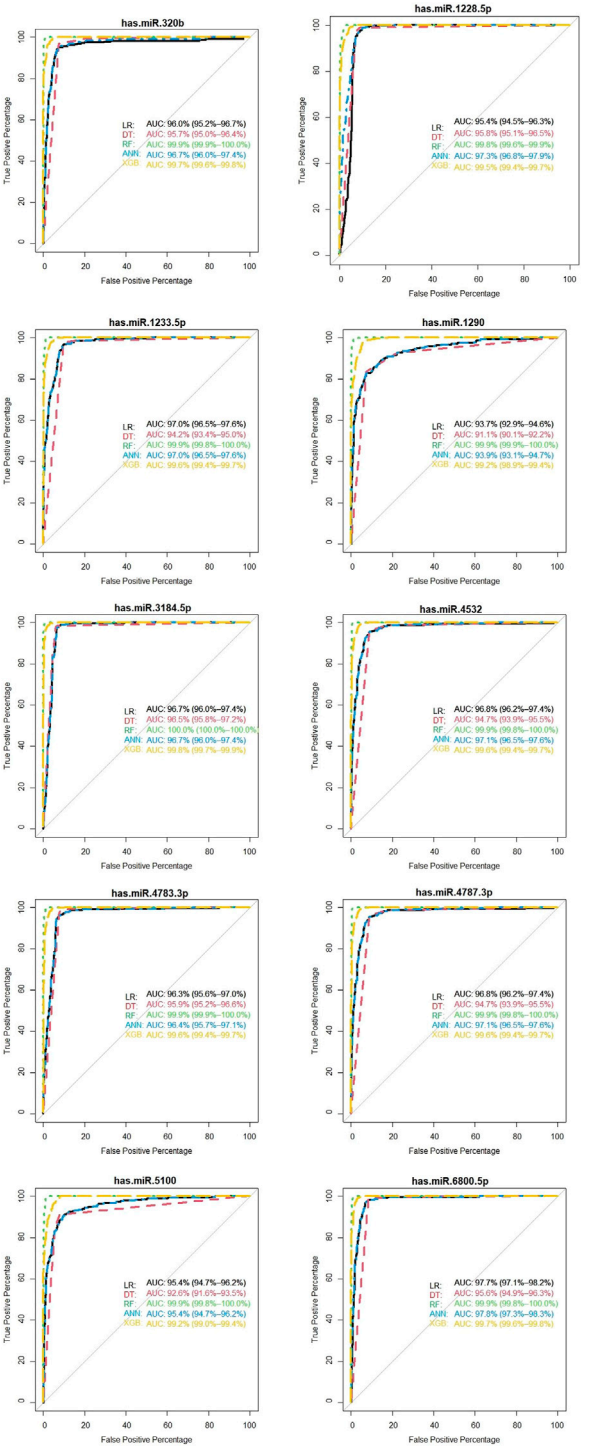

FIGURE 4 | Diagnostic performance of the 10 identified serum miRNA signatures in the internal (GSE106817) data.

area under the receiver operating characteristic (AUC), positive predictive value, negative predictive values, and Kappa (Collins et al., 2015). The ROC curves were analyzed by "pROC" in the R software.

Further, two online tools are applied to assess the biological plausibility of the selected miRNAs. To compare the microarray expression profiles of ovarian cancer to the non-cancer group, GEO2R is an interactive web tool that allows users to compare two or more groups of samples in a GEO Series. This procedure will enable the users to identify indicators that are differentially expressed across experimental conditions. To do this end, the limma R package implemented in GEO2R online tool, which generated adjusted *p*-value, B-statistic (or log-odds), Log2-fold change (logfc), and moderated t-statistic. MiRNet is an online tool for precision miRNA and xeno-miRNA analysis and functional interpretation. This tool contains a large amount of high-quality scientific data that connects miRNAs to their targets and other associated compounds (Fan et al., 2016).

RESULTS

GSE106817 included 2,568 miRNAs. Of those, LASSO and ENET identified 76 and 162 miRNAs, respectively. Then, the dataset was divided with a ratio of 70:30 for the training and testing set, respectively. For the training set, there were 2,156 samples and there were 923 samples in the testing set. The training set had 224 ovarian cancerous and 1,932 non-cancerous samples. After balancing the training data, the samples of non-cancerous decreased to 1,121 and cancerous samples increased to 1,035. Model fitting and tuning parameter selection by 5-fold cross-validation were done on the training data. The dataset with reduced features is classified using LR (statistical), DT and RF (tree-based), ANN and XGB (machine learning) classifier. In this study, the features with higher importance (over 80%) implemented in proposed models are shown in **Table 2**.

We identified 10 potential miRNAs hsa-miR-5100, hsa-miR-6800-5p, hsa-miR-1233-5p, hsa-miR-4532, hsa-miR-4783-3p, hsa-miR-4787-3p, hsa-miR-1228-5p, hsa-miR-1290, hsa-miR-3184-5p, and hsa-miR-320b from the GSE106817 datasets and were defined as the candidate miRNAs for ovarian cancer diagnosis. It is clear that hsa-miR-1233-5p, hsa-miR-4783-3p, hsa-miR-5100, and hsa-miR-1290 are features identified by both feature selection methods. hsa-miR-320b and hsa-miR-6800-5p have been identified as important features by LASSO, and hsa-miR-4532, hsa-miR-3184-5p, hsa-miR-4787-3p, and hsa-miR-1228-5p have been recognized by ENET.

The results of GEO2R (generated by the limma) are presented in Table function (**Table 2**). Note that the column of adjusted *p*-value is generally recommended as the primary statistic in the interpretation of results. The miRNAs with the smallest *p*-values will be the most reliable, and column B shows that the represented miRNAs are differentially expressed and logfc presented change between normal and cancerous conditions. As shown in **Table 2**, all upregulated miRNAs have logfc > 2 and all of miRNAs have adjusted *p*-value <0.0001. Based on the 10 selected miRNAs, the final machine

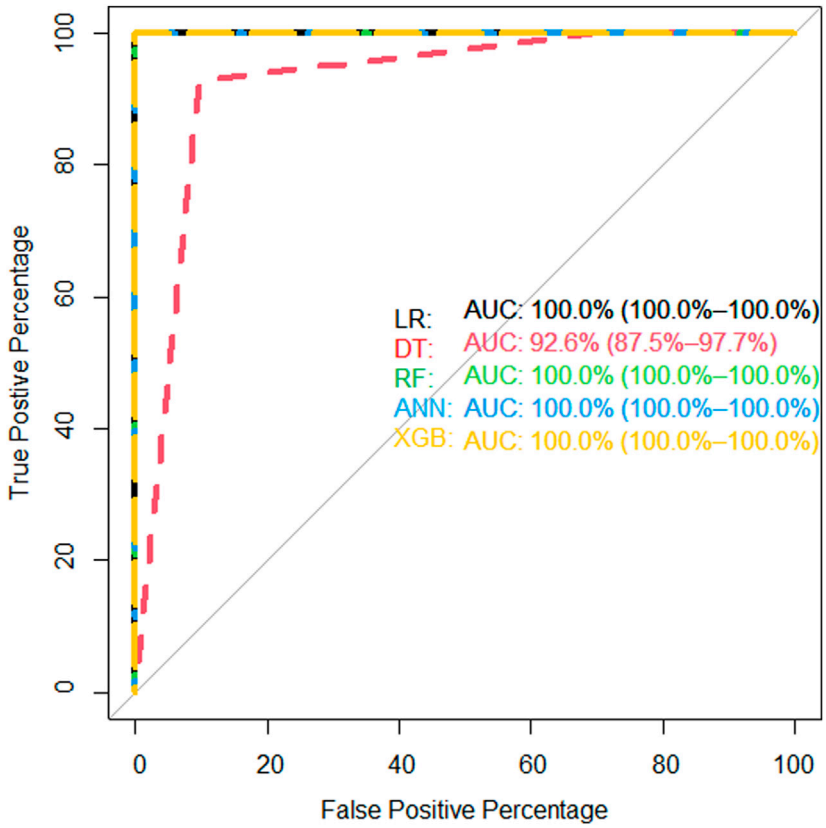

FIGURE 5 | AUC of proposed models of all identified microRNAs in the internal (GSE106817) validation data.

learning models with optimal hyperparameters are presented in **Table 3**.

We showed the expression levels of these 10 identified miRNAs in the internal datasets using a boxplot (**Figure 2**); among them, seven miRNAs (hsa-miR-320b, hsa-miR-5100, hsa-miR-4783-3p, hsa-miR-1290, hsa-miR-4532, hsa-miR-4787-3p, and hsa-miR-1233-5p) identified the most significantly up-regulated in ovarian cancer samples compared to non-cancer samples. The heatmap using the "pheatmap" package shows differences between samples in each group. In **Figure 3** (the heatmap of GSE106817), the miRNAs has-mir-3184-5p, has-mir-6800-5p, and has-mir-1228-5p in the left hand side of the figure show a significantly low expression level in the ovarian cancer group (red color). However, hsa-mir-5100, hsa-mir-1290, hsa-mir-320b, hsa-mir-1233-5p, hsa-mir-4532, hsa-mir-4783-3p, and hsa-mir-4783-3p have the high expression levels in the cancerous group (light yellow color). The individual AUCs of these 10 identified miRNAs are listed in **Figure 4** which shows that each of 10 miRNAs has high AUC in all proposed models. Next, AUCs of all selected miRNAs are presented in **Figure 5** which clearly indicates that all moles, except DT, have above 99% AUC. All miRNA-target gene interactions are represented in **Figure 6**. The purple circles represent the target genes implicated in cancer-related pathways that are shown by yellow circles.

Model Evaluation in External Validation Data

Given the robust performance of 10 miRNAs in the internal datasets, we further examined their performance in independent external validation (GSE113486). External validation dataset (GSE113486) has 40 ovarian cancer patients and 52 non-cancer controls (43% ovarian cancer, 57% non-cancer). We found that all the miRNAs had high performance and could efficiently distinguish the ovarian cancer samples from non-cancer controls.

As shown in **Figure 7**, hsa-miR-320b, hsa-miR-1233-5p, hsa-miR-3184-5p, and hsa-miR-4783-3p have 100% of AUC in all proposed models. In the external validation dataset (GSE113486), the AUC of each candidate miRNAs was over 95% (minimum AUC: 95.7%, maximum AUC: 100%) for ovarian cancer classification (**Figure 7**). From **Supplementary Figure S2**, it is clear that, except DT, other machine learning models have an AUC over 100% in the external validation dataset with 10 selected miRNAs.

The models that yielded the highest AUC, accuracy, and sensitivity are shown in **Table 3**. As displayed in **Table 3** (and also **Supplementary Figure S2**), we found four models yielded 100% AUC; however, DT did not have a strong performance because it is weak learner (Drucker and Cortes, 1996).

Finally, to make use of our prediction models, the practitioners can give the values of the 10 selected miRNAs in the online excel

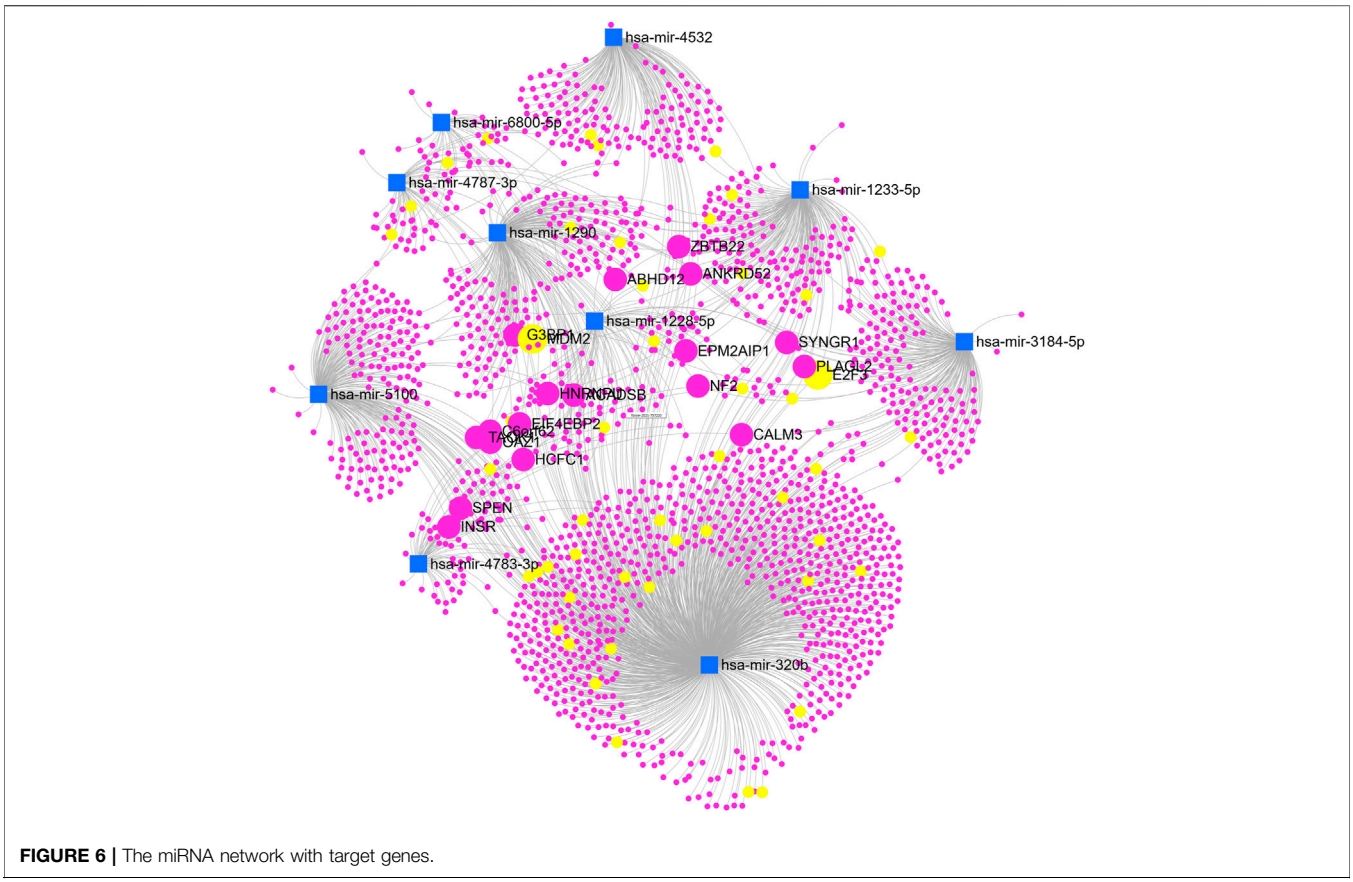

FIGURE 6 | The miRNA network with target genes.

sheet (https://ufile.io/t2exrfph) and calculate the probability of the ovarian cancer for the patient (**Supplementary Figure S1**).

DISCUSSION

In the early phases, ovarian cancer is mostly asymptomatic or existent with only non-specific symptoms (Desai et al., 2014; Tuncer et al., 2020). Intervention at this phase makes ovarian cancer almost curable, and thus, early detection and diagnosis are critical to decrease the incidence and mortality of ovarian cancer (Zhang et al., 2011). Therefore, in this study, we used effective strategies and identified 10 miRNAs (hsa-miR-5100, hsa-miR-6800-5p, hsa-miR-1233-5p, hsa-miR-4532, hsa-miR-4783-3p, hsa-miR-4787-3p, hsa-miR-1228-5p, hsa-miR-1290, hsa-miR-3184-5p, and hsa-miR-320b) as strong potential biomarkers for ovarian cancer. We found that these miRNAs (all together) had high enough prediction accuracy for identification of ovarian cancer from non-cancer (logistic regression had an AUC 100%, sensitivity 100%, and specificity 100%; decision trees had an AUC 92.60%, sensitivity 92.5%, and specificity 90.38%; random forest had an AUC 100%, sensitivity 95%, and specificity 100%; artificial neural network had an AUC 100%, sensitivity 100%, and specificity 100.0%; and XGBoost had an AUC 100%, sensitivity 97.50%, and specificity 100%). Furthermore, hsa-miR-5100, hsa-miR.4532, hsa-miR.4783.3p, and hsa-miR-320b were more stable in the discovery and validation datasets.

Biological Insight

There is evidence in the literature for the biomarkers included in our study. Huang et al. (2011) showed that modulation of miR-5100 could potentially be employed as a therapeutic target for cancer (Huang H. et al., 2015). It has shown that major target gene of miR-5100 is AZIN1. AZIN1 gene encodes antizyme inhibitor 1, the first member of this gene family that is ubiquitously expressed, and is localized in the nucleus and cytoplasm. Overexpression of antizyme inhibitor one gene has been associated with increased proliferation, cellular transformation, and tumorigenesis (Hu et al., 2017). Also, our result is important about the relationship between ovarian cancer and miR-5100 because of target gene function. Tuncer et al. (2020) suggested that hsa-miR-6800-5p is an effective biomarker for ovarian cancer. MiR-1233 is considered an oncomiRNA since it targets p53, inhibiting its function in RCC (Iwamoto et al., 2014). Hu et al., (2017) showed that miR-4532 is involved in the multidrug resistance formation in breast cancer by targeting hypermethylated cancer 1 (*HIC-1*), a tumor-suppressor gene (Feng et al., 2018). Also, hsa-miR-4783-3p has a major target of INSM1/IA-1 (insulinoma-associated one gene) (http://mirdb.org/) and this gene is a developmentally regulated zinc-finger transcription factor, exclusively expressed in the foetal pancreas and nervous systems, and in tumours of neuroendocrine origin (Juhlin et al., 2020). Li et al., 2016 suggest that miRNA-1228 is deregulated, and the most encompassed biological pathways are apoptosis-related (Li et al., 2016). In another study, miR-1290 is

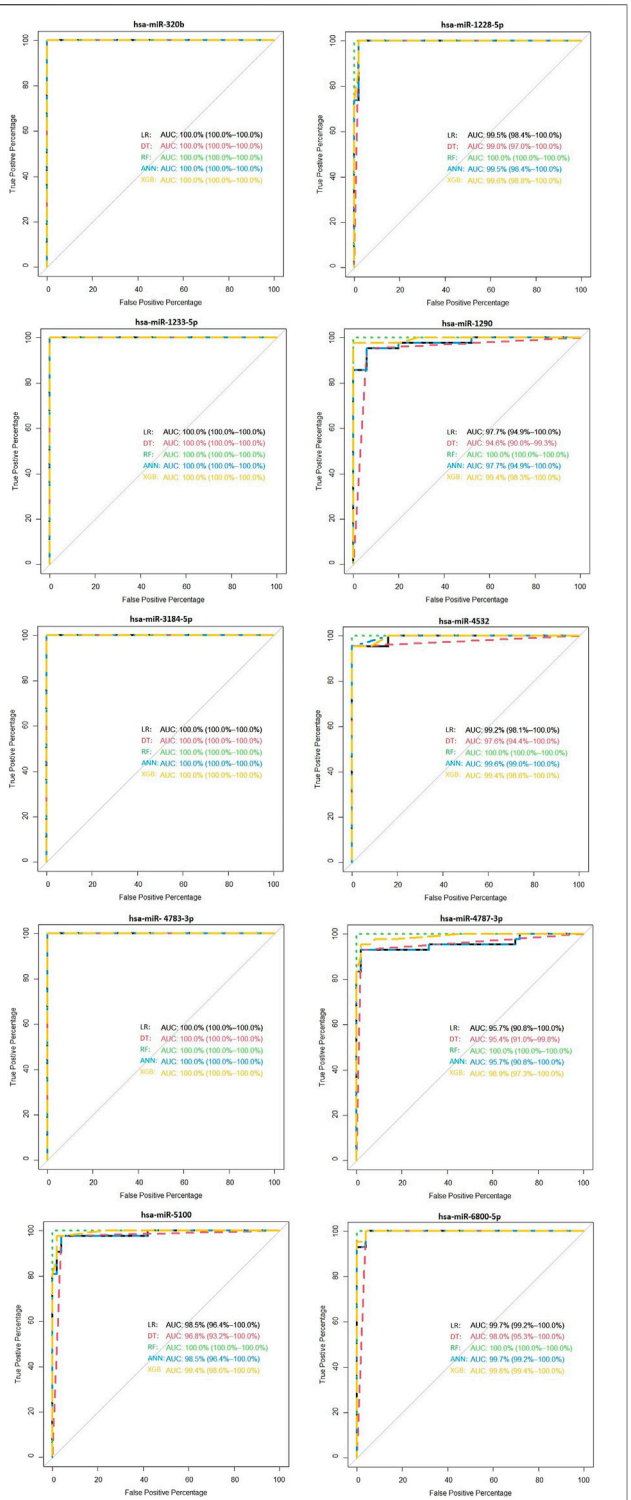

FIGURE 7 | Validation of the diagnostic performance of the three selected miRNA high signatures in the external (GSE113486) data.

significantly overexpressed in patients with high-grade serous ovarian carcinoma (HGSOC) and they suggested that it is a new potential diagnostic biomarker for HGSOC. Exosomal miR-1290 is a potential biomarker of high-grade serious ovarian carcinoma (Cortez et al., 2018). The study of Tuncer et al. (2020) revealed that miR-320b belonged to the miR-320 family which has low expression levels in ovarian cancer. Prior studies indicated that decreased expression level of the miR-320 family is associated to activate cell proliferation (Tuncer et al., 2020). We have analyzed the major target genes of the upregulated miRNA interactions (**Supplementary Figure S3**). We found only two gene interactions with string database system, especially TP53 and HIC1 genes associated with a related system in human metabolism (**Supplementary Figure S3**).

Strengths and Limitations

This study has several strengths. Firstly, we applied logistic regression and four of the main machine learning approaches to predict ovarian cancer. Secondly, we identified predictive models to predict the ovarian cancer. Our findings provided strong evidence that the serum miRNA profile represented a promising diagnostic biomarker for ovarian cancer. Thirdly, we used two robust variable selection approaches to identify the important miRNAs. Finally, we evaluated the prediction accuracy of the proposed prediction models in both internal and external data to provide more robust results for practical and clinical applications.

However, there were certain limitations in our study. We had relatively small sample size in ovarian cancer group. Other limitations were the pathological information such as the tumor stage, age, or other factors which were not available in GSE106817 dataset. Nonetheless, the prediction accuracy of our model has high enough (100% AUC) for clinical use. But we still suggest further study to consider age, stage, and other unrecognized factors associated with ovarian cancer that has not included in the current paper. Also, we restricted our analysis to ovarian cancer patients and non-cancer controls, and we did not evaluate the capability of these miRNAs to distinguish ovarian cancer from other cancers.

CONCLUSION

In this paper, we used the state-of-the-art machine learning algorithms along with so-called penalized statistical approaches to model ovarian cancer with miRNA data. Our algorithms selected 10 important miRNA that can predict the ovarian cancer with an AUC of 100%. Our findings provided significant evidence that the serum miRNA profile represents a promising diagnostic biomarker for ovarian cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RA, NG, and FH contributed to the conception and design of the study. RA, NG, and FH performed the statistical analysis.

FH wrote the first draft of the manuscript. TE wrote the biological discussion section. RA, NG, PS, TE, FH and PS wrote sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

This study was supported by Tabriz University of Medical Sciences with grant number 66567.

ACKNOWLEDGMENTS

The authors would like to thank all those who spent their valuable time participating in this research project, and we are also immensely grateful to the reviewers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.724785/full#supplementary-material

REFERENCES

Abdulqader, Q. M. (2017). Applying the Binary Logistic Regression Analysis on the Medical Data. *Sci. J. Univ. Zakho* 5 (4), 330–334. doi:10.25271/2017.5.4.388

Asano, N., Matsuzaki, J., Ichikawa, M., Kawauchi, J., Takizawa, S., Aoki, Y., et al. (2019). A Serum microRNA Classifier for the Diagnosis of Sarcomas of Various Histological Subtypes. *Nat. Commun.* 10 (1), 1299–1310. doi:10.1038/s41467-019-09143-8

Banka, H., and Dara, S. (Editors) (2012). *Feature Selection and Classification for Gene Expression Data Using Evolutionary Computation*. Vienna, Austria: 23rd International Workshop on Database and Expert Systems Applications, IEEE.

Chen, T., and Guestrin, C. (Editors) (2016). "Xgboost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., and Cho, H. (2015). Xgboost: Extreme Gradient Boosting. *R. Package Version 04-2* 1 (4), 1–4. doi:10.1038/onc.2014.117

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. M. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). *Circulation* 131 (2), 211–219. doi:10.1161/circulationaha.114.014508

Cortez, A. J., Tudrej, P., Kujawa, K. A., and Lisowska, K. M. (2018). Advances in Ovarian Cancer Therapy. *Cancer Chemother. Pharmacol.* 81 (1), 17–38. doi:10.1007/s00280-017-3501-8

Deb, B., Uddin, A., and Chakraborty, S. (2018). miRNAs and Ovarian Cancer: An Overview. *J. Cel Physiol* 233 (5), 3846–3854. doi:10.1002/jcp.26095

DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., et al. (2018). A Review of Machine Learning in Obesity. *Obes. Rev.* 19 (5), 668–685. doi:10.1111/obr.12667

Desai, A., Xu, J., Aysola, K., Qin, Y., Okoli, C., Hariprasad, R., et al. (2014). Epithelial Ovarian Cancer: An Overview. *World J. Transl. Med.* 3 (1), 1. doi:10.5528/wjtm.v3.i1.1

Drotár, P., Gazda, J., and Smékal, Z. (2015). An Experimental Comparison of Feature Selection Methods on Two-Class Biomedical Datasets. *Comput. Biol. Med.* 66, 1–10. doi:10.1016/j.compbiomed.2015.08.010

Drucker, H., and Cortes, C. (1996).Boosting Decision Trees *Adv. Neural Inf. Process. Syst.*, 479–485.

Falzone, L., Candido, S., Salemi, R., Basile, M. S., Scalisi, A., McCubrey, J. A., et al. (2016). Computational Identification of microRNAs Associated to Both Epithelial to Mesenchymal Transition and NGAL/MMP-9 Pathways in Bladder Cancer. *Oncotarget* 7 (45), 72758–72766. doi:10.18632/oncotarget.11805

Falzone, L., Lupo, G., Rosa, G. R. M., Crimi, S., Anfuso, C. D., Salemi, R., et al. (2019). Identification of Novel MicroRNAs and Their Diagnostic and Prognostic Significance in Oral Cancer. *Cancers* 11 (5), 610. doi:10.3390/cancers11050610

Fan, Y., Siklenka, K., Arora, S. K., Ribeiro, P., Kimmins, S., and Xia, J. (2016). miRNet - Dissecting miRNA-Target Interactions and Functional Associations through Network-Based Visual Analysis. *Nucleic Acids Res.* 44 (W1), W135–W141. doi:10.1093/nar/gkw288

Feng, F., Zhu, X., Wang, C., Chen, L., Cao, W., Liu, Y., et al. (2018). Downregulation of Hypermethylated in Cancer-1 by miR-4532 Promotes Adriamycin Resistance in Breast Cancer Cells. *Cancer Cell Int.* 18 (1), 127–212. doi:10.1186/s12935-018-0616-x

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33 (1), 1–22. doi:10.18637/jss.v033.i01

Greenlee, R. T., Hill-Harmon, M. B., Murray, T., and Thun, M. (2001). Cancer Statistics, 2001. *CA Cancer J. Clin.* 51 (1), 15–36. doi:10.3322/canjclin.51.1.15

Harter, P., Reuss, A., Pfisterer, J., Pujade-Lauraine, E., Ray, I., and du Bois, A. (2008).The Role of Surgical Outcome as Prognostic Factor in Advanced Epithelial Ovarian Cancer. A Project of the AGO-OVAR and GINECO–Prognostic Factor Surgical Outcome in Advanced Ovarian Cancer. *Geburtshilfe Frauenheilkd.* 68, 1–4. doi:10.1055/s-0028-1088605

Hassanipour, S., Ghaem, H., Arab-Zozani, M., Seif, M., Fararouei, M., Abdzadeh, E., et al. (2019). Comparison of Artificial Neural Network and Logistic Regression Models for Prediction of Outcomes in Trauma Patients: A Systematic Review and Meta-Analysis. *Injury* 50 (2), 244–250. doi:10.1016/j.injury.2019.01.007

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin, Germany: Springer Science & Business Media.

Hu, X., Chen, J., Shi, X., Feng, F., Lau, K. W., Chen, Y., et al. (2017). RNA Editing of AZIN1 Induces the Malignant Progression of Non-small-cell Lung Cancers. *Tumour Biol.* 39 (8), 1010428317700001. doi:10.1177/1010428317700001

HuangHailiJiang, Y., Wang, Y., Chen, T., Yang, L., et al. (2011). miR-5100 promotes tumor growth in lung cancer by targeting Rab6. *Cancer letters* 362 (1), 15–24. doi:10.1016/j.canlet.2015.03.004

Huang, H., Jiang, Y., Wang, Y., Chen, T., Yang, L., He, H., et al. (2015). miR-5100 Promotes Tumor Growth in Lung Cancer by Targeting Rab6. *Cancer Lett.* 362 (1), 15–24. doi:10.1016/j.canlet.2015.03.004

Huang, J., Li, Y.-F., and Xie, M. (2015). An Empirical Analysis of Data Preprocessing for Machine Learning-Based Software Cost Estimation. *Inf. Softw. Tech.* 67, 108–127. doi:10.1016/j.infsof.2015.07.004

Iorio, M. V., Visone, R., Di Leva, G., Donati, V., Petrocca, F., Casalini, P., et al. (2007). MicroRNA Signatures in Human Ovarian Cancer. *Cancer Res.* 67 (18), 8699–8707. doi:10.1158/0008-5472.can-07-1936

Iwamoto, H., Kanda, Y., Sejima, T., Osaki, M., Okada, F., and Takenaka, A. (2014). Serum miR-210 as a Potential Biomarker of Early clear Cell Renal Cell Carcinoma. *Int. J. Oncol.* 44 (1), 53–58. doi:10.3892/ijo.2013.2169

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Juhlin, C. C., Zedenius, J., and Höög, A. (2020). Clinical Routine Application of the Second-Generation Neuroendocrine Markers ISL1, INSM1, and Secretagogin in Neuroendocrine Neoplasia: Staining Outcomes and Potential Clues for Determining Tumor Origin. *Endocr. Pathol.* 31 (4), 401–410. doi:10.1007/s12022-020-09645-y

Lee, Y. S., and Dutta, A. (2009). MicroRNAs in Cancer. *Annu. Rev. Pathol. Mech. Dis.* 4, 199–227. doi:10.1146/annurev.pathol.4.110807.092222

Lheureux, S., Gourley, C., Vergote, I., and Oza, A. M. (2019). Epithelial Ovarian Cancer. *Lancet* 393 (10177), 1240–1253. doi:10.1016/s0140-6736(18)32552-2

Li, X., Ding, Z., Zhang, C., Zhang, X., Meng, Q., Wu, S., et al. (2016). MicroRNA-1228* Inhibit Apoptosis in A549 Cells Exposed to fine Particulate Matter. *Environ. Sci. Pollut. Res.* 23 (10), 10103–10113. doi:10.1007/s11356-016-6253-9

Liang, X., and Jacobucci, R. (2020). Regularized Structural Equation Modeling to Detect Measurement Bias: Evaluation of Lasso, Adaptive Lasso, and Elastic Net. *Struct. Equ. Modeling* 27 (5), 722–734. doi:10.1080/10705511.2019.1693273

Lin, X.-J., Chong, Y., Guo, Z.-W., Xie, C., Yang, X.-J., Zhang, Q., et al. (2015). A Serum microRNA Classifier for Early Detection of Hepatocellular Carcinoma: a Multicentre, Retrospective, Longitudinal Biomarker Identification Study with a Nested Case-Control Study. *Lancet Oncol.* 16 (7), 804–815. doi:10.1016/s1470-2045(15)00048-0

Lisboa, P. J., and Taktak, A. F. G. (2006). The Use of Artificial Neural Networks in Decision Support in Cancer: a Systematic Review. *Neural Netw.* 19 (4), 408–415. doi:10.1016/j.neunet.2005.10.007

Lunardon, N., Menardi, G., and Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *R J.* 6 (1). doi:10.32614/rj-2014-008

Menard, S. (2010). *Logistic Regression: From Introductory to Advanced Concepts and Applications*. Thousand Oaks, CA, California: Sage.

Nam, E. J., Yoon, H., Kim, S. W., Kim, H., Kim, Y. T., Kim, J. H., et al. (2008). MicroRNA Expression Profiles in Serous Ovarian Carcinoma. *Clin. Cancer Res.* 14 (9), 2690–2695. doi:10.1158/1078-0432.ccr-07-1731

Okun, O., and Priisalu, H. (Editors) (2007). "Random forest for Gene Expression Based Cancer Classification: Overlooked Issues," in *Iberian Conference on Pattern Recognition and Image Analysis* (Springer).

Qi, Y. (2012). "Random Forest for Bioinformatics," in *Ensemble Machine Learning* (Springer), 307–323. doi:10.1007/978-1-4419-9326-7_11

Reid, B. M., Permuth, J. B., and Sellers, T. A. (2017). Epidemiology of Ovarian Cancer: a Review. *Cancer Biol. Med.* 14 (1), 9–32. doi:10.20892/j.issn.2095-3941.2016.0084

Sherriff, A., Ott, J., and Team, A. S. (2004). Artificial Neural Networks as Statistical Tools in Epidemiological Studies: Analysis of Risk Factors for Early Infant Wheeze. *Paediatr. Perinat Epidemiol.* 18 (6), 456–463. doi:10.1111/j.1365-3016.2004.00592.x

Stoltzfus, J. C. (2011). Logistic Regression: a Brief Primer. *Acad. Emerg. Med.* 18 (10), 1099–1104. doi:10.1111/j.1553-2712.2011.01185.x

Tuncer, S. B., Erdogan, O. S., Erciyas, S. K., Saral, M. A., Celik, B., Odemis, D. A., et al. (2020). miRNA Expression Profile Changes in the Peripheral Blood of Monozygotic Discordant Twins for Epithelial Ovarian Carcinoma: Potential New Biomarkers for Early Diagnosis and Prognosis of Ovarian Carcinoma. *J. Ovarian Res.* 13 (1), 99–15. doi:10.1186/s13048-020-00706-8

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., and Blockeel, H. (2008). Decision Trees for Hierarchical Multi-Label Classification. *Mach Learn.* 73 (2), 185–214. doi:10.1007/s10994-008-5077-3

Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F. X., et al. (2005). Gene Selection from Microarray Data for Cancer Classification-A Machine Learning Approach. *Comput. Biol. Chem.* 29 (1), 37–46. doi:10.1016/j.compbiolchem.2004.11.001

Yao, Y., Ding, Y., Bai, Y., Zhou, Q., Lee, H., Li, X., et al. (2020). Identification of Serum Circulating MicroRNAs as Novel Diagnostic Biomarkers of Gastric Cancer. *Front. Genet.* 11, 591515. doi:10.3389/fgene.2020.591515

Zhang, B., Cai, F. F., and Zhong, X. Y. (2011). An Overview of Biomarkers for the Ovarian Cancer Diagnosis. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 158 (2), 119–123. doi:10.1016/j.ejogrb.2011.04.023

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. B* 67 (2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 November 2021
doi: 10.3389/fgene.2021.763590

Identification of Immune-Related Genes Associated With Bladder Cancer Based on Immunological Characteristics and Their Correlation With the Prognosis

Zhen Kang[1,2], *Wei Li*[1,2], *Yan-Hong Yu*[1,2], *Meng Che*[1], *Mao-Lin Yang*[1,2], *Jin-Jun Len*[1,2], *Yue-Rong Wu*[1] *and Jun-Feng Yang*[1,2]*

[1]*The Affiliated Hospital, Kunming University of Science and Technology, Kunming, China,* [2]*Department of Urology, The First People's Hospital of Yunnan Province, Kunming, China*

OPEN ACCESS

Edited by:
Miguel E. Renteria,
QIMR Berghofer Medical Research Institute, Australia

Reviewed by:
Luis M. García-Marín,
QIMR Berghofer Medical Research Institute, Australia
Cen Wu,
Kansas State University, United States

***Correspondence:**
Jun-Feng Yang
yjfkmmc@163.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *24 August 2021*
Accepted: *08 November 2021*
Published: *26 November 2021*

Citation:
Kang Z, Li W, Yu Y-H, Che M, Yang M-L, Len J-J, Wu Y-R and Yang J-F (2021) Identification of Immune-Related Genes Associated With Bladder Cancer Based on Immunological Characteristics and Their Correlation With the Prognosis. Front. Genet. 12:763590. doi: 10.3389/fgene.2021.763590

Background:To identify the immune-related genes of bladder cancer (BLCA) based on immunological characteristics and explore their correlation with the prognosis. Methods: We downloaded the gene and clinical data of BLCA from the Cancer Genome Atlas (TCGA) as the training group, and obtained immune-related genes from the Immport database. We downloaded GSE31684 and GSE39281 from the Gene Expression Omnibus (GEO) as the external validation group. R (version 4.0.5) and Perl were used to analyze all data. Result: Univariate Cox regression analysis and Lasso regression analysis revealed that 9 prognosis-related immunity genes (PIMGs) of differentially expressed immune genes (DEIGs) were significantly associated with the survival of BLCA patients ($p < 0.01$), of which 5 genes, including *NPR2, PDGFRA, VIM, RBP1, RBP1* and *TNC*, increased the risk of the prognosis, while the rest, including *CD3D, GNLY, LCK,* and *ZAP70,* decreased the risk of the prognosis. Then, we used these genes to establish a prognostic model. We drew receiver operator characteristic (ROC) curves in the training group, and estimated the area under the curve (AUC) of 1-, 3- and 5-year survival for this model, which were 0.688, 0.719, and 0.706, respectively. The accuracy of the prognostic model was verified by the calibration chart. Combining clinical factors, we established a nomogram. The ROC curve in the external validation group showed that the nomogram had a good predictive ability for the survival rate, with a high accuracy, and the AUC values of 1-, 3-, and 5-year survival were 0.744, 0.770, and 0.782, respectively. The calibration chart indicated that the nomogram performed similarly with the ideal model. Conclusion:We had identified nine genes, including *PDGFRA, VIM, RBP1, RBP1*, *TNC*, *CD3D, GNLY, LCK,* and *ZAP70*, which played important roles in the occurrence and development of BLCA. The prognostic model based on these genes had good accuracy in predicting the OS of patients and might be promising candidates of therapeutic targets. This study may provide a new insight for the diagnosis, treatment and prognosis of BLCA from the perspective of immunology. However, further experimental studies are necessary to reveal the underlying mechanisms by which these genes mediate the progression of BLCA.

Keywords: bladder cancer, ssGSEA, tumor immunity, immune characteristics, urology

INTRODUCTION

Bladder cancer (BLCA) is one of the 10 most common cancers around the world, with 550,000 new cases and 200,000 deaths in 2018 (Richters et al., 2020). The risk of BLCA is 1 in 74 for men and 1 in 301 for women, and in the past decade, the number of new cases of BLCA has increased by 32% (Fitzmaurice et al., 2019). As we all know, non-muscle invasive bladder cancer (NMIBC) and muscle invasive bladder cancer (MIBC) are the two main types of bladder cancer. When patients progress from NMIBC to MIBC, their overall survival (OS) rate significantly decreases (Cao et al., 2020a; Tran et al., 2021), and about one-third of NMIBC patients will develop MIBC (Sylvester et al., 2006). As we all know, bladder cancer diagnosis represents a challenge for clinicians, and currently available diagnostic and staging tools include: 1) urine cytological analysis; 2) cystoscopy and pathological biopsy; 3) computed tomography or magnetic resonance imaging. However, all of the above-mentioned tools have some defects, such as low sensitivity or demands for invasive operation (van Rhijn et al., 2009). Tumor markers, as a new research tool, can not only help clinicians understand the characteristics of tumors, but also help early diagnosis, improve prognosis and carry out risk stratification and targeted therapy for tumor patients (Bratu et al., 2021). So far, there have been many studies on blood (Dohn et al., 2021), tissue and urine markers (Aibara et al., 2021; Chen et al., 2021; Tosev et al., 2021) of bladder cancer, and clinical guidelines are paying more attention to the application of clinical tumor markers (Witjes et al., 2021). Especially, genetic testing often performs better in predicting the prognosis, and multi-gene prognostic models are gradually becoming the choice of more clinicians (Qu et al., 2021).

In recent years, immune checkpoint inhibitors (ICPIs) have revolutionized the treatment paradigm for most malignant tumors with persistent positive responses even observed in advanced and refractory cancers (Bindal et al., 2021). Therefore, exploring the interaction between tumor cells and immunity can help clinicians gain a deeper understanding of the occurrence, development and metastasis of BLCA (Guan et al., 2021). So far, a lot of recent studies have performed the analysis of the immune characteristics of BLCA patients, which have fully demonstrated that immune genes have higher predictive values of the prognosis, and provide better clinical guidance than routine clinical features or risk models (Cao et al., 2020b; Wang et al., 2021; Zhang et al., 2021). However, these studies only evaluated the immunological characteristics of BLCA from the view of immune cell infiltration, and lacked the exploration on the tumor-immune interaction and its potential values of predicting the prognosis of BLCA.

The tumor microenvironment (TME) consists of immune cells, stromal cells, extracellular vesicles and other molecules. A study showed that TME was an important regulator of gene expression and was closely involved in the occurrence, development and treatment of tumors (Kumari et al., 2021). The immune system and immune response play a crucial role in TME (Dzobo, 2020). In this study, we innovatively used single-sample gene set enrichment analysis (ssGSEA) to classify BLCA patients into a high-immune (Immunity_H) group and a low-immune (Immunity_L) group, and then explored the tumor-immune interaction, related molecular characteristics, and the potential prognosis from the perspective of immune-difference-related genes. Finally, we used these genes and the machine learning method of the Least Absolute Shrinkage and Selection Operator (Lasso) algorithm to establish a prognostic model, and validated the stability and repeatability of the model in an external independent data set.

MATERIALS AND METHODS

Data Collection

The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) is a landmark cancer genomics program that molecularly describes over 20,000 primary cancer, and matches normal samples spanning 33 cancer types. This joint effort between National Cancer Institute (NCI) and the National Human Genome Research Institute began in 2006, and has produced over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already led to improvements in our ability to diagnose, treat, and prevent cancer, will remain publicly available for anyone in the research community to use. We downloaded FPKM standardized RNA-seq data, clinical information and tumor mutation burden (TMB) information from the TCGA-BLCA cohort in TCGA database.

ImmPort (https://www.immport.org/) is funded by the National Institute of Health (NIH) and National Institute of Allergy and Infectious Diseases (NIAID) in support of the NIH mission to share data with the public. We clicked the "Resources" button on the Immport database homepage, then clicked the "Gene Lists" button on the "Resources" page, and finally clicked the "Gene Summary" to download immune-related genes.

Gene Expression Omnibus (GEO) is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. We downloaded two data sets (GSE31684 and GSE39281) recording bladder cancer transcriptome genes (RNA-seq) and clinical information in the GEO database. After processing the data with Perl, we obtained two gene expression matrices. Then, we used the "sva" package in the R language (version 4.0.5) to merges the two expression matrices and eliminate batch effects.

Data Analysis

(A) TMB analysis: we used BLCA mutation data in the TCGA database and Perl language to calculate the number of base mutations in each BLCA sample. (B) Single-sample gene set enrichment analysis (ssGSEA) and hierarchical cluster analysis: we used R packages (GSVA, GSEABase and limma) to perform ssGSEA to calculate the immune score of each BLCA sample according to 29 immune gene sets composed of different types of immune cells with different functions, pathways and checkpoints (Alhamdoosh et al., 2017). Firstly, the rank of gene expression values in a given BLCA sample was normalized, and then the enrichment

score (ES) was calculated using the empirical cumulative distribution function. Each ssGSEA score XI was converted to XI′ by bias normalization to obtain the scores of different immune cells and immune-related functions in each sample. Then, we used the hierarchical clustering method of Euclidean distance and Ward linkage to do the immune stratification of BLCA patients. Meanwhile, we also made use of the T-distribution stochastic neighbor embedding (tSNE) algorithm to determine the immune stratification of BLCA patients through RtSEN package (Gardner et al., 2021). (C) Evaluation of tumor immune microenvironment: based on ESTIMATE algorithm, BLCA transcriptome data was utilized to predict stromal cell score, immune cell score and tumor purity, and then the content of these two types of cells was predicted, from which StromalScore, ImmuneScore and EstimateScore were determined (Yoshihara et al., 2013). (D) Tumor-infiltrating immune cells analysis: CIBERSORT, an R tool, was used for the deconvolution of the expression matrix of human immune cell subtypes according to linear support vector regression. This method is based on a known reference set and provides a set of gene expression characteristics of 22 immune cell subtypes. Therefore, we used the CIBERSORT method to do the calculation for the abundance of infiltrating immune cells in BLCA samples (Newman et al., 2015). (E) Immune differential genes determining the immune stratification: the limma package was utilized to select differentially expressed genes (DEGs) among people with different immune stratification (| log2 fold change | > 1.50 and FDR <0.05), and then we obtained immune-related genes from ImmPort (Bhattacharya et al., 2014). DEIGs were obtained through the intersection of immune genes and DEGs. (E) Prognostic markers: the survival package was utilized to do the univariate Cox regression analysis ($p < 0.05$) to identify the markers of significant prognosis-related immunity genes (PIMGs).

Gene Set Pathway Enrichment Analysis

Gene set enrichment analysis was performed via the GSEA software (version 4.1.0) to analyze TCGA-BLCA transcriptomes for the identification of the key signaling pathways involved in DEGs.

The major Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways involved in the up-regulation of the Immunity_H and Immunity_L subgroups ($p < 0.05$, FDR <0.01) were selected. R (version 4.0.5) was used to perform further analysis, and visualize the results. Then, we obtained transcription factors associated with the occurrence and development of bladder cancer from the CISTROME project, extracted differentially expressed transcription factors (DETFs) from the total DEGs, and used Pearson correlation coefficient analysis to construct the regulatory network of PIMGs and DETFs (R > 0.3 and FDR <0.01) (Mei et al., 2017). Finally, the protein-protein interaction (PPI) network analysis was performed using STRING (String-db.org/).

Constructing and Validating the Prognostic Model of the Immune-Related Genes

We used the LASSO Cox regression model in R package (Dalal et al., 2012) "glmnet" to find genes significantly associated with the prognosis to construct the prognostic model of BLCA (PMB). The risk score was calculated as the following formula: $riskScore = \sum_{i=1}^{9} \beta_i * LPIMG_i$, where $LPIMG_i$ represented the i-th LPIMG (Lasso-prognosis-related immunity genes), and $\boldsymbol{\beta}_i$ represented the expression coefficient of $LPIMG_i$ obtained from Lasso regression analysis. All cases were classified into a low-risk group and a high-risk group based on the median risk score, and we performed the Kaplan-Meier survival analysis to compare the survival status between the high-risk group and the low-risk group. In order to verify the predictive power of PMB, the receiver operator characteristic (ROC) curve was drawn to calculate the area under the curve (AUC) of 1-, 3-, and 5-year survival. We conducted Kaplan-meier, logarithmic rank, ROC curve and calibration analysis using "timeROC," "rms," "survival," and "survminer" software packages in R language. Based on the risk score calculated by PMB, Pearson correlation coefficient, Spearman correlation coefficient and corrplot package were used to evaluate the correlation between the risk score and overall survival, immune cell infiltration, immune checkpoint molecules and TMB. $p < 0.05$ of the critical value for the significant correlation was set. Eventually, univariate and multivariate Cox regression analysis of the risk scores of the constructed PMB and patients' clinical characteristics (age, sex, stage) was performed to verify the accuracy of the independence of PMB-based risk characteristics. Based on the above factors, we created a nomogram using the R packages of "rms", "nomogramEX" and "regplot." Finally, the ROC and calibration chart were drew to determine the suitability of our established nomogram for potential clinical applications.

RESULTS

Identification of Two Subtypes of BLCA Using Immune Analysis

In order to fully evaluate the immunological characteristics of BLCA, we used the ssGSEA to analyze 414 tumor samples from the TCGA-BLCA cohort. According to the ssGSEA scores and hierarchical clustering method, BLCA cases were divided into two clusters. The average score of the immune microenvironment of the first cluster was 0.62, and the average score of the immune microenvironment of the second cluster was 0.49. Thus, the first cluster was set as the Immunity_H (high) group, and the second cluster as the Immunity_L (low) group (**Figures 1A,B**). The tSNE was further used to analyze the immune levels for different BLCA patients and the same classification was obtained (**Supplementary Figure S1A**). The results of ESTIMATE analysis indicated that EstimateScore (419.27 ± 1649.47), ImmuneScore (750.39 ± 886.17), and StromalScore (-331.12 ± 910.28) in the Immunity_H group were significantly higher than those which were (−2283.37 ± 727.70), (−620.62 ± 352.59), and (−1662.75 ± 487.21), respectively, in the Immunity_L group (Wilcox test, $p < 0.001$) (**Figure 1C**). CIBERSORT was used to detect the degree of immune cell infiltration in the tumor, which found that the differences between the Immunity_H group and the Immunity_L group in T cells CD4 naive, T cells CD4 memory resting, T cells

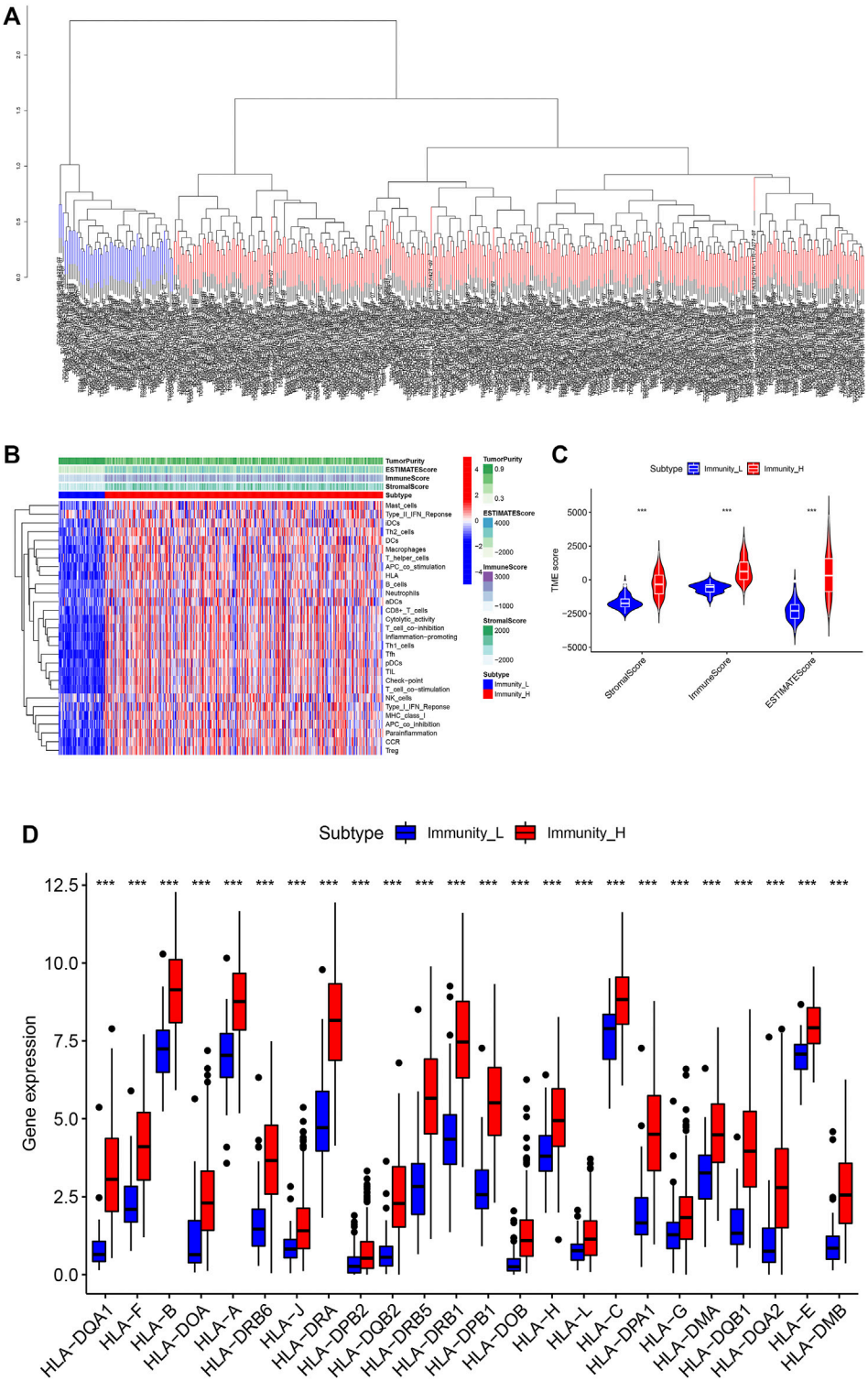

FIGURE 1 | (A) The two immune types of BLCA patients, the red part was the high immune group, the blue was the low immune group. **(B)** The status of immune infiltration and tumor microenvironment (TME) in the TCGA-BLCA cases. **(C)** The comparisons of StromalScore, ESTIMATEScore, and ImmuneScore between the two subtypes. **(D)** The comparison of expression level of HLA gene between the two subtypes. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

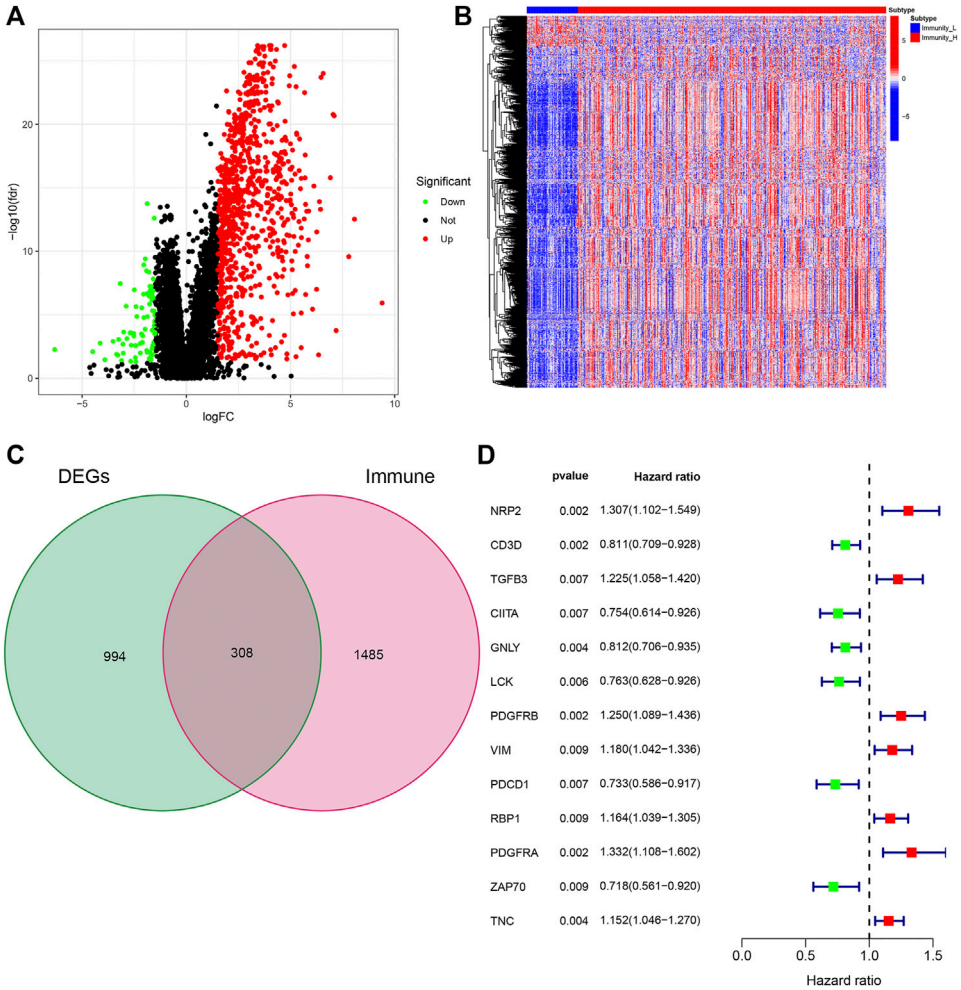

FIGURE 2 | (A) Volcano plot of all differentially expressed genes (DEGs) showing the log2 (fold change) and FDR value of each gene. **(B)** DEGs expression between the two subtypes in the heat maps. **(C)**Veen plot based on the intersection of DEGs and human immune genes. **(D)** Forest plot based on univariable Cox proportional hazards regression analysis showing the prognosis-related immunity genes (PIMGs) and their hazard ratios.

CD4 memory activated, NK cells resting, NK cells activated, Macrophages M1 and Mast cells activated were significant (**Supplementary Figure S1B**). The expression of human leukocyte antigen (HLA) genes in the both groups was examined, which suggested that most of HLA genes significantly increased in the Immunity_H group and significantly decreased in the Immunity_L group (Wilcox test, $p < 0.05$) (**Figure 1D**). Based on our results, we believed that immune response might play important roles in the development of BLCA.

Identification of immune-related genes associated with bladder cancer and their correlation with prognosis.

We further studied the expression of differential genes of immune stratification in BLCA patients. The FDR values and log2 fold change multiples of the immune differential genes in the Immunity_H group and the Immunity_L group were showed in **Figure 2A**. After primarily screening, we totally identify 994 DEGs, of which 812 genes were up-regulated and 82 genes were down-regulated (**Figure 2B**). Subsequently, 308 DEGs were selected as DEIG using the ImmPort database (**Figure 2C**). Univariate Cox regression analysis indicated that 13 PIMGs had significant association with the survival of BLCA patients in DEIGs ($p < 0.01$), of which seven genes, including *NPR2, TGFB3, PDGFRB, PDGFRA, VIM, RBP1, RBP1* and *TNC*, increased the risk of prognosis, while the rest, including *CD3D, CIITA, GNLY, LCK, PDCD1* and *ZAP70,* were conducive to survival (**Figure 2D**).

Identifying Prognosis-Related Genes and Constructing the Prognostic Model

LASSO Cox regression analysis was performed on 13 selected PIMGs (**Figures 3A,B**). Finally, 9 LPIMGs were identified and their risk-correlation coefficients were calculated to determine the prognosis of BLCA patients. The risk score was calculated as follows: riskScore = NRP2*0.0101119 + CD3D*−0.1990949 + GNLY*−0.1241769 + LCK*−0.0519549 + VIM*0.1464182 + RBP1*0.1038418 + PDGFRA*0.1589969 + ZAP70*−0.12644895 + TNC*0.0693184. Data from TCGA

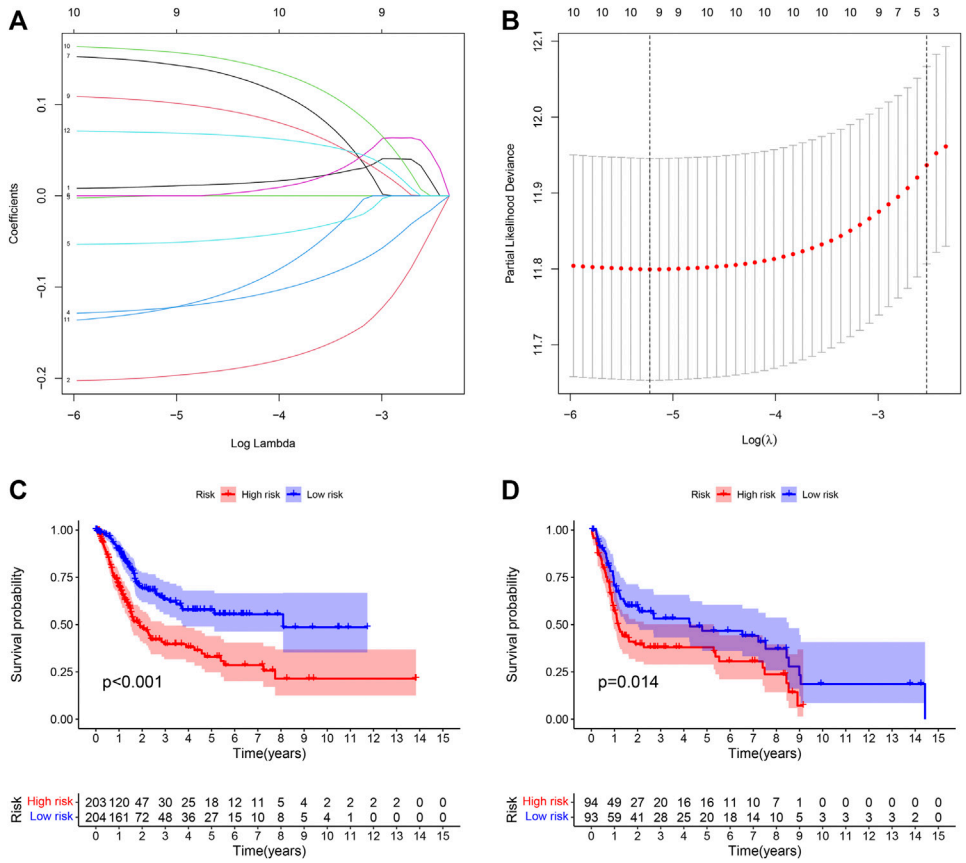

FIGURE 3 | (A) LASSO coefficient curves were selected with simulation parameters set to 1000. **(B)** 10-fold cross-validation of selecting tuning parameter in the LASSO model. **(C)** Kaplan-Meier survival analysis of the PMB-based risk signature in the TCGA-BLCA cohort. **(D)** Kaplan-Meier survival analysis of the PMB-based risk signature in the GSECD cohort.

was selected as the training group, the risk score of each BLCA case in this group was calculated, and all cases were classified into the high-risk group (203 patients) and the low-risk group (204 patients) based on the median risk score of 0.4886 (**Supplementry Data S1**; **Supplementary Figure S1C**). The correlation analysis indicated that the risk score had significant negative correlation with the survival time of BLCA patients which gradually decreased with the increase of the risk score (**Supplementary Figures S1D,E**). The Kaplan-meier curve showed that the difference in overall survival (OS) between the high-risk group and the low-risk group was significant, and patients in the low-risk group had a longer overall survival time than those in the high-risk group (Log-rank test, $p < 0.0001$) (**Figure 3C**). In order to evaluate the predictive power and accuracy of PMB-based risk characteristics, the ROC curves of the training group were drawn, and the AUC values of 1-, 3- and 5-year survival were 0.688, 0.719, and 0.706, respectively (**Supplementary Figure S1F**). The accuracy of the prognostic model was verified by the calibration chart, which suggested that the predicted value of the prognostic model was in good consistence with the actual value (**Figure 3H**). Besides, GSE31684 and GSE39281 were used as the external validation group, and we combined their data (GSECD) using R "sva" package to further confirm the accuracy and feasibility of the prognostic model, and the number of deaths in the high-risk group increased significantly (**Supplementary Data S2**). Then, the Pierce correlation analysis and Kaplan-Meier curves suggested that the constructed PMB-based risk characteristics still had good predictive power in the external validation group (**Figure 3D**).

Combined Analysis of Tumor Immune Microenvironment and the Model of Prognosis

In order to investigate the correlation between immunotherapy and bladder cancer, 14 immune checkpoint inhibitors inlcuding *BTLA, GITR, TNFRSF14, IDO, LAG-3, PD-1, PD-L1, PD-L2, CD28, CD40, CD80, CD137, CD27*, and *Ctla-4* were selected for analysis. It was found that the risk score had negative correlation with the BTLA, CD27, CD40, CD80, and TNFRSF14 expression, which had significant differences in different risk groups (**Supplementary Figure S2**), indicating that tumor

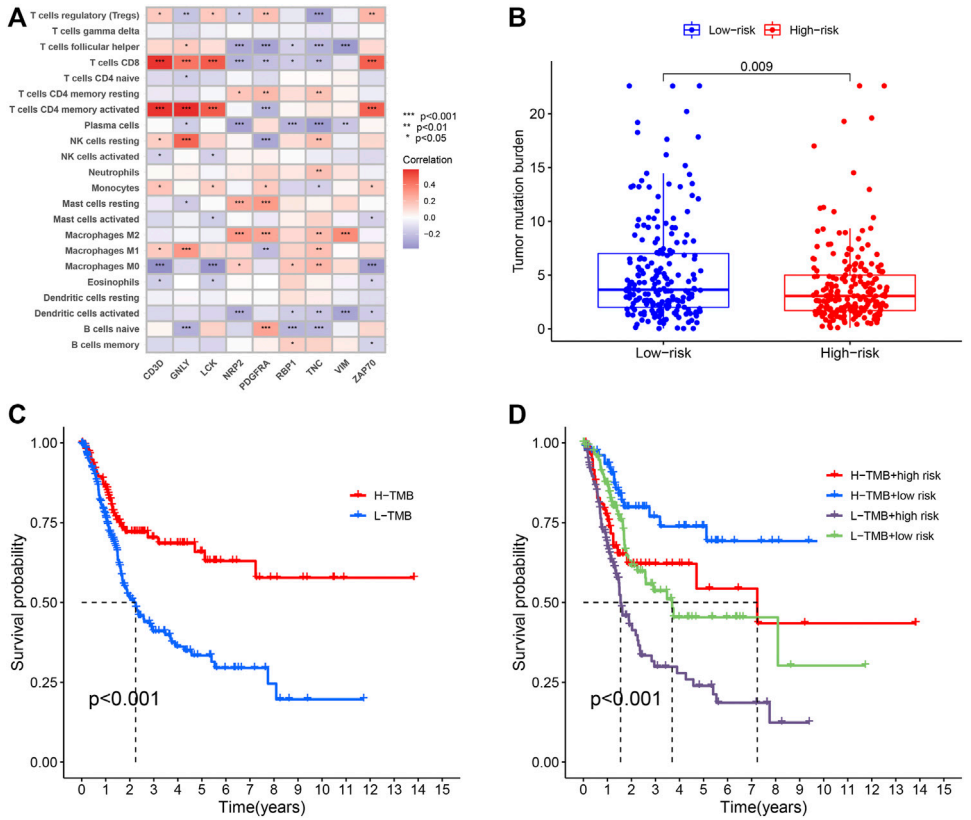

FIGURE 4 | (A) Heatmap showing the correlation of the prognostic model of BLCA (PMB)-based risk signature with immune cell infiltration. The red suggesting the positive correlation while blue suggesting the negative correlation. **(B)** The comparison of TMB between High- and low-risk groups. **(C)** Kaplan-Meier survival analysis of the TMB in the TCGA-BLCA cohort. **(D)** Kaplan-Meier survival analysis of four groups stratified by combining the TMB and the PMB-based risk signature in the TCGA-BLCA cohort.

immunosuppression might lead to an increased risk score of patients. TYK2 and ACE2 were also differentially expressed in different risk groups, and with the increase of the risk score, their expression decreased (**Figure 4A**). In the TCGA-BLCA cohort, the TMB of patients in the high-risk group was significantly lower than that in the low-risk group (p = 0.009) (**Figure 4B**). In order to find the potential correlation between TMB and the prognosis of patients, according to the TMB cutoff value of 4.632, we divided the patients into the high TMB group, and the low TMB group (**Supplementary Data S3**). We found that the survival time of patients in the high TMB group was significantly lower than that of the low TMB group ($p < 0.001$) (**Figure 4C**). In order to evaluate the outcomes of patients more comprehensively, we investigated whether the combination of the risk score and TMB could be a more accurate prognostic marker. We integrated PMB-based risk characteristics with TMB, stratified all samples into the H-TMB/high risk, H-TMB/low risk, L-TMB/high risk, and L-TMB/low risk groups. **Figure 4D** suggested that differences between groups were significant (log-rank test, $p < 0.0001$), and in the H-TMB/low risk group, patients had the longest overall survival. The above results together suggested that the risk score had positive correlation with the degree of malignant tumor.

Establishing a Nomogram With Clinical Features

Due to the significant correlation between the risk score and the degree of malignant tumor, univariate and multivariate Cox regression analysis for age, sex, and stage as covariates was conducted to test the potential possibility of the risk score as an independent prognostic factor for BLCA patients, of which the results showed that the PMB based risk characteristics had a p value less than 0.001, confirming that the PMB based risk characteristics could be used to predict the prognosis of BLCA patients (**Table 1**). Combined with the above factors, we constructed a nomogram (**Figure 5A**) to expand the clinical application and usability of PMB. The total score of each patient was obtained by calculating and summing the score for each prognostic parameter. The higher the total score was, the worse the patient's clinical outcome was. The ROC curve showed that the nomogram had a good predictive ability for the survival rate, with a high accuracy, and the AUC values of 1-, 3-, and 5-year survival were 0.744, 0.770, and 0.782, respectively (**Figure 5B**). In addition,

TABLE 1 | Univariable and multivariable Cox analysis of clinical characteristics and riskScore in the TCGA-BLCA cohort.

	Univariate cox regression				Multivariate cox regression			
ID	HR	HR.95L	HR.95H	pvalue	HR	HR.95L	HR.95H	pvalue
Age	1.039588391	1.022252149	1.057218636	6.04E-06	1.035655214	1.018448898	1.053152224	4.16E-05
Gender	0.913510834	0.6440517	1.29570661	0.611966738	0.870137802	0.611038106	1.239104053	0.440546516
Stage	1.822621822	1.479575308	2.245205288	1.68E-08	1.545728603	1.243151661	1.921951269	8.92E-05
riskScore	3.002207633	2.158508821	4.175683964	6.55E-11	2.483209078	1.749461945	3.524699316	3.58E-07

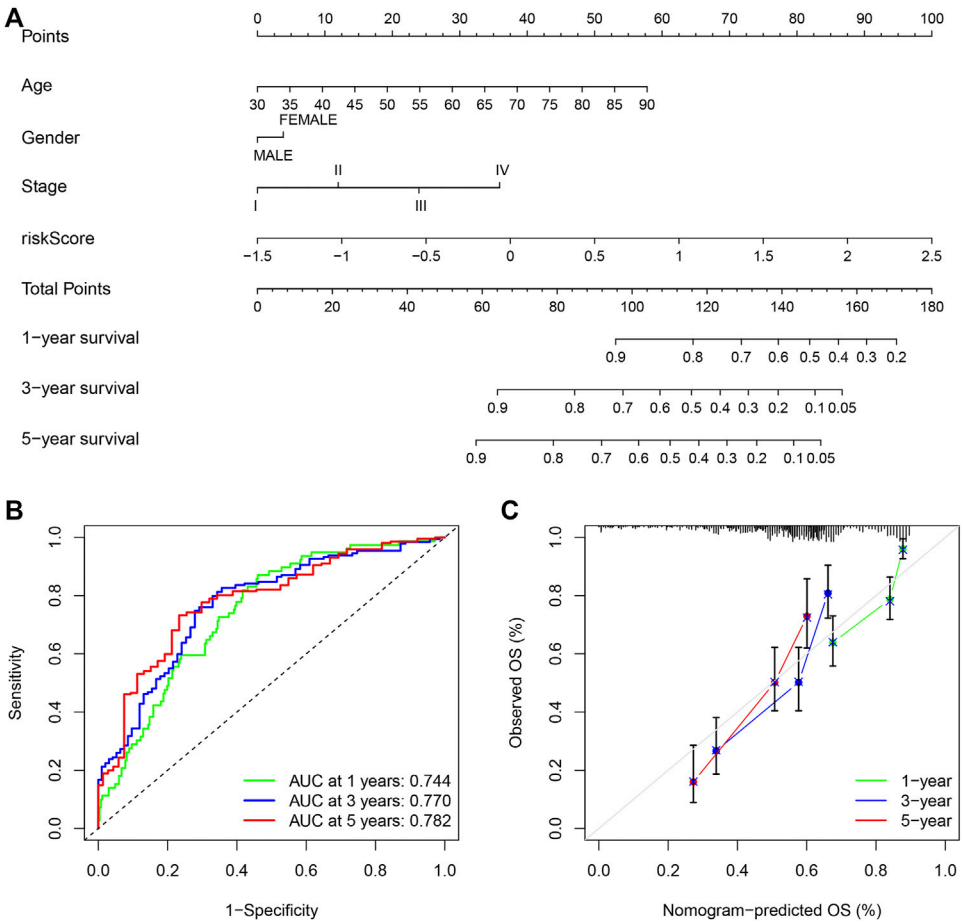

FIGURE 5 | (A) Nomogram of age, gender, stage and risk score as independent prognostic factors for predicting overall survival. **(B)** The receiver operator characteristic (ROC) curves and the area under the curve (AUC) of the predictions for 1-, 3-, and 5-years of the nomogram for TCGA-BLCA cohort. **(C)** The calibration chart of the nomogram for TCGA-BLCA cohort.

the calibration chart indicated that the nomogram performed similarly with the ideal model (**Figure 5C**).

Gene Set Pathway Enrichment Analysis

GSEA revealed that immune-associated pathways in the Immunity_H group were highly active, including the signaling pathway of T cell receptor, the pathway of antigen processing and presentation, cytokine involved immune response, and hematopoietic cell lineage. Additionally, various pathways of immune-associated disease were identified in the Immunity_H roup, including asthma, primary immune deficiency, graft-versus-host disease, allograft rejection, thyroid disease related to autoimmune, and immunity to leishmania infection (**Figure 6A**). In order to clarify the role of the multi-dimensional regulatory network of immune molecules in the occurrence and development of bladder cancer, we firstly explored the upstream mechanism of PIMG. By combining differential expression analysis with data from the CISTROME database, we identified transcription factors significantly associated with the BLCA prognosis. For the Immunity_H subtype, a total of 7 up-regulated transcription factors were identified. **Figure 6B** showed the regulatory network of BLCA TF-PIMGs. PPI analysis was further conducted and we confirmed the significant correlation between BLCA TF and PIMG (**Figure 6C**).

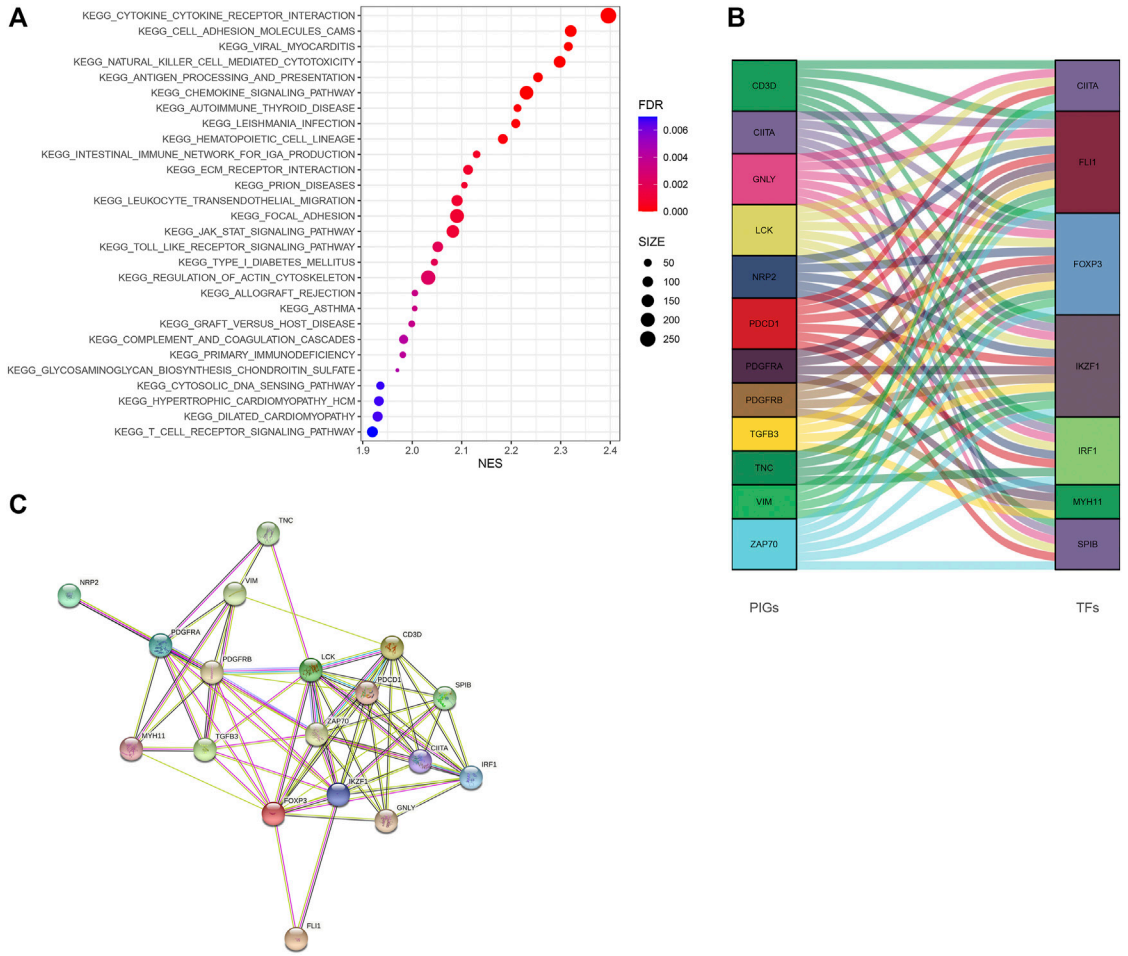

FIGURE 6 | (A) The Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of DEGs. **(B)** Alluvial diagram of the BLCA TFs and PIMGs revealing their regulatory network. **(C)** PPI network between BLCA TFs and PIMGs. Network nodes representing proteins, and edges representing protein-protein associations, including both functional and physical protein associations. Line thickness indicating the strength of data support. The thicker line representing the higher confidence. KEGG, Kyoto Encyclopedia of Genes and Genomes; BLCA TF, Bladder cancer transcription factors; PIMGs, prognosis- associated immunity genes.

DISCUSSION

In our study, we collected gene expression data and clinical information of BLCA from the public databases. A total of 9 immune-related prognostic genes were identified by the Lasso analysis. Subsequently, a nine-gene prognostic model of BLCA (PMB) was established. We integrated clinical characteristics and risk scores to establish a nomogram. The ROC curve and calibration chart verified the prognostic accuracy of the nomogram. The high-risk KEGG analysis showed that the main functions of genes in the high-risk group were closely related to immunity. Finally, TMB had a significant correlation with the prognosis of patients, and had a potential connection with the PMB model. These findings strongly implied that immunity played a non-negligible role in the occurrence of BLCA.

We used Lasso regression to establish a PMB model, and used the file **Supplementary Data S4, S5** and the code **Supplementary Data S6** to achieve the repetition of the results of the model. Wu and Ma (2015) believed that in the researches of genetic analysis, most of the analyzed genes were expected to be "noise", and only a few were related to the results and phenotypes. In the process of eliminating "noise" genes, a variety of machine learning methods (LASSO, adaptive LASSO, SCAD, and MCP) had been used. For the low-dimensional genomics data, stable approaches were widely developed, while for the high-dimensional genomics data, the development of approaches was limited. Therefore, in the process of screening genes, a variety of machine learning methods are worthy of our further trial and comparison. Ren et al. (Ren et al., 2019) believed that because gene expression might show heavy tailed distributions (especially for the high-expression genes), or be contaminated, the gene regulation relationship inference based on non-robust methods might be biased. Thus, we proposed a robust network based on the regularization and variable selection method for high-dimensional genomics data in cancer prognosis, and correspondingly also used "regnet" package in R language. The robust and regularized AFT model was fitted by the network

penalty, and 9 prognostic genes were obtained by Lasso regression analysis. As we deeply understand the new machine learning methods, we will introduce new methods such as "regnet" at the design stage of bioinformatics analysis to explore more possibilities in the future.

In the past, some prognostic models of BLCA patients had been established (Dong et al., 2021; Liu et al., 2021), but in these studies, the tumor-immune-TMB interaction have not been fully considered. For the TCGA-BLCA patients, we firstly, based on immunogenomics analysis, divided the patients into the high immune (Immunity_H) subtype and the low immune (Immunity_L) subtype. Compared with the Immunity_L subtype, we found that the Immunity_H subtype showed stronger immune cell infiltration and higher expression of HLA genes, which suggested stronger immunogenicity. The Immunity_H subtype had abundant immune-related characteristics, and was rich in a lot of cancer-related pathways, such as leukemia, pancreatic cancer, and melanoma. What's more, the results of our study found the potential association between immune activity and pathway activity for BLCA patients.

According to the expression of these 9 immune genes, the PMB based risk characteristics was developed, as a new predictive tool for the prognosis of BLCA, and was validated in the two data sets of GSE31684 and GSE39281. The results showed that the OS curves of patients with high- and low-risk scores were significantly different. Based on the risk characteristics of PMB combined with immune invasion, the prognosis of patients was predicted, and the survival time of patients in the low-risk/Immune-L group was the longest. Of the 9 genes used to construct the PMB, five oncogenes, namely *NRP2, VIM, RBP1, PDGFRA*, and *TNC*, were promising therapeutic targets. *NRP2* (Neuropilin 2) can regulate the activity of vascular endothelial growth factor-activated receptor, protein binding, and heparin binding, and take part in the positive regulation of angiogenesis, endothelial cell proliferation, cell adhesion, endothelial cell migration and other pathways, and its targeted drugs can treat hypoplasia in children (Estrada et al., 2021). *VIM* (Vimentin) is involved in the combination of double-stranded RNA, the formation of cytoskeleton, the formation of the lens of the eye, negative regulation of neuron projection development, astrocyte development, and cytokine-mediated signaling pathway. *RBP1* (Retinol-binding protein 1) is involved in several physiological functions (Gao et al., 2020), including regulation of metabolism and retinol transport. *PDGFRA* (platelet derived growth factor receptor alpha) mutations cause a variety of heterogeneous gastrointestinal mesenchymal tumors (Ricci et al., 2015), and *TKIs* inhibiting the most common driving mutations in *KIT* or *PDGFRA* might have brought about radical changes in treating gastrointestinal stromal tumors in the past 20 years (Zalcberg, 2021). *TNC* (enascin-C) is a large extracellular matrix glycoprotein that promotes cell adhesion and tissue remodeling, and is involved in the transduction of cellular signaling pathways (Spenlé et al., 2021). These findings encourage us to explore the molecular mechanisms of these genes in BLCA in the future.

It has been proved that immune checkpoint inhibitors, such as nivolumab, pembrolizumab, ipilimumab, atezolizumab, avelumab, and durvalumab, are effective for treating metastatic urological neoplasms (Petzold et al., 2021). We found that five immune checkpoint inhibitors, including CD27, CD40, CD80, BTLA, and TNFRSF14, were significantly negatively correlated with the risk score of patients, indicating that the risk of patients would increase with the increase of immune expression. Sensitivity to CD40 ligation-induced apoptosis might be a new mechanism to eliminate tumor transformation of urothelial cells. The important adaptive mechanism for the occurrence and development of transitional cell carcinoma might be CD40 expression loss (Bugajska et al., 2002). CD80 is an essential membrane antigen for the activation of T lymphocytes. CD80 monoclonal antibody inhibits the adjuvant stimulation of CD80, and prevents the differentiation of B lymphocytes into plasma cells, which plays a prominent role in the treatment of tumors (Vackova et al., 2021). CD27 and CD40 belong to the tumor necrosis factor receptor (TNFR) family. As a co-stimulatory pathway molecule, CD40 has been proven to be very successful in combination with pro-active drug antibody targets in both single-dose therapy and combination therapy (Peters et al., 2009). CD27 can stimulate the anti-tumor effect of monoclonal antibodies, and the stimulation of CD27 on the T cells surface and NK cells can increase the release of chemokines (Seidel et al., 2016). B- and T-lymphocyte attenuator (BTLA) is also known as B- and T-lymphocyte-associated protein. Under normal physiological conditions, the combination of BTLA and its ligand HVEM can inhibit the over-activation of lymphocytes *in vivo*, and prevent the immune system from damaging itself (Yu et al., 2021). Finally, TNFRSF14 might exert a tumor suppressor effect in bladder cancer by inducing cell apoptosis and inhibiting proliferation (Zhu and Lu, 2018). These immune-related studies are worthy of further exploration in the immunotherapy of bladder cancer in the future.

BLCA patients with a higher level of TMB had better prognosis, and when TMB increased, the response rate of immunotherapy was higher, implying that TMB might be an independent biomarker that can provide the guidance for more effective immunotherapy and improve the prognosis of BLCA (Ready et al., 2019). In addition, we observed that PMB was significantly correlated with TMB. Compared the AUC values of the ROC curves between the two groups, the combination of TMB and PMB also could predict the survival of patients. These findings suggested that risk characteristics based on PMB might help measure the responses to immunotherapy.

There are some limitation in our study. Firstly, the underlying mechanism of how the identified 9 LPIMGs regulate the BLCA process is still unclear, and their biological functions need to be further explored by experiments. Secondly, the development and verification of this model are only based on the public databases, and thus more clinical research data is still necessary to verify its effectiveness. Lastly, regarding the machine learning methods, we used Lasso regression to perform the gene screening and completed all the research, but Lasso regression may not be the most ideal method to identify relevant features (such as

gene expression). The new method of "regnet" are worthy to use in the future study.

CONCLUSION

In summary, we had identified nine genes, including *PDGFRA, VIM, RBP1, RBP1*, *TNC*, *CD3D, GNLY, LCK,* and *ZAP70*, which played important roles in the occurrence and development of BLCA. The prognostic model based on these genes had good accuracy in predicting the OS of patients and might be promising candidates of therapeutic targets. In addition, further experimental studies are necessary to reveal the underlying mechanisms by which these genes mediate the progression of BLCA.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

ZK and J-FY, conceiving and designing the study; ZK, WL, Y-HY, MC, M-LY, J-JL, and Y-RW, collecting the data; ZK, WL, Y-HY, MC, M-LY, J-JL, and Y-RW, analyzing and interpreting the data; ZK, writing the manuscript; J-FY, providing critical revisions that are important for the intellectual content; ZK, WL, Y-HY, MC, M-LY, J-JL, Y-RW, and J-FY, approving the final version of the manuscript.

FUNDING

This research was supported by "Yunnan Health Training Project of High Level Talents" (H-2017046).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.763590/full#supplementary-material

Supplementry Figure 1 | (A) Validation of immunophenotype via tSNE; **(B)** The comparison of the immune cell infiltration level between two subtypes; **(C,E)** Distributions of the risk score, survival time, and survival status in the TCGA-BLCA cohort; **(D)** Correlation analysis of the risk score and survival time in the TCGA-BLCA cohort; **(F)** The ROC curves and AUC of the predictions for 1, 3, and 5 years of the PMB-based risk signature for TCGA-BLCA cohort.

Supplementary Figure 2 | Boxplots and Scatter plots depicting correlation between the PMB-based risk signature and gene expression of immune checkpoint inhibitors. **(A,B)** BTLA; **(C,D)** CD27; **(E,F)** CD40; **(G,H)** CD80; **(I,J)** TNFRSF14.

Supplementary Data 1 | Distributions of the risk score, survival time, and survival status in the TCGA-BLCA cohort.

Supplementary Data 2 | Gene expression, risk score and risk stratification of patients in the GEO database.

Supplementary Data 3 | Dividing TCGA-BLCA patients into different groups based on TMB

Supplementary Data 4 | Gene expression significantly associated with the patient's prognosis in the TCGA-BLCA cohort Supplement data

Supplementary Data 5 | The expression of genes significantly associated with the patient's prognosis in the GEO database.

Supplementary Data 6 | The code of Lasso regression analysis (R 4.0.5).

REFERENCES

Aibara, N., Miyata, Y., Araki, K., Sagara, Y., Mitsunari, K., Matsuo, T., et al. (2021). Detection of Novel Urine Markers Using Immune Complexome Analysis in Bladder Cancer Patients: A Preliminary Study. *In Vivo* 35 (4), 2073–2080. doi:10.21873/invivo.12476

Alhamdoosh, M., Ng, M., Wilson, N. J., Sheridan, J. M., Huynh, H., Wilson, M. J., et al. (2017). Combining Multiple Tools Outperforms Individual Methods in Gene Set Enrichment Analyses. *Bioinformatics* 33 (3), btw623–424. doi:10.1093/bioinformatics/btw623

Bhattacharya, S., Andorf, S., Gomes, L., Dunn, P., Schaefer, H., Pontius, J., et al. (2014). ImmPort: Disseminating Data to the Public for the Future of Immunology. *Immunol. Res.* 58 (2-3), 234–239. doi:10.1007/s12026-014-8516-1

Bindal, P., Gray, J. E., Boyle, T. A., Florou, V., and Puri, S. (2021). Biomarkers of Therapeutic Response with Immune Checkpoint Inhibitors. *Ann. Transl Med.* 9 (12), 1040. doi:10.21037/atm-20-6396

Bratu, O., Marcu, D., Anghel, R., Spinu, D., Iorga, L., Balescu, I., et al. (2021). Tumoral Markers in Bladder Cancer (Review). *Exp. Ther. Med.* 22 (1), 773. doi:10.3892/etm.2021.10205

Bugajska, U., Georgopoulos, N. T., Southgate, J., Johnson, P. W., Graber, P., Gordon, J., et al. (2002). The Effects of Malignant Transformation on Susceptibility of Human Urothelial Cells to CD40-Mediated Apoptosis. *J. Natl. Cancer Inst.* 94 (18), 1381–1395. doi:10.1093/jnci/94.18.1381

Cao, R., Yuan, L., Ma, B., Wang, G., and Tian, Y. (2020). Immune-related Long Non-coding RNA Signature Identified Prognosis and Immunotherapeutic Efficiency in Bladder Cancer (BLCA). *Cancer Cell Int* 20, 276. doi:10.1186/s12935-020-01362-0

Cao, Y., Tian, T., Li, W., Xu, H., Zhan, C., Wu, X., et al. (2020). Long Non-coding RNA in Bladder Cancer. *Clinica Chim. Acta* 503, 113–121. doi:10.1016/j.cca.2020.01.008

Chen, H., Liu, Y., Cao, C., Xi, H., Chen, W., Zheng, W., et al. (2021). CYR61 as a Potential Biomarker for the Preoperative Identification of Muscle-Invasive Bladder Cancers. *Ann. Transl Med.* 9 (9), 761. doi:10.21037/atm-19-4511

Dalal, S. R., Shekelle, P. G., Hempel, S., Newberry, S. J., Motala, A., and Shetty, K. D. (2012). *A Pilot Study Using Machine Learning and Domain Knowledge to Facilitate Comparative Effectiveness Review Updating*. Report No.: 12-EHC069-EF. Rockville (MD): Agency for Healthcare Research and Quality.

Dohn, L. H., Thind, P., Salling, L., Lindberg, H., Oersted, S., Christensen, I. J., et al. (2021). Circulating Forms of Urokinase-type Plasminogen Activator Receptor in Plasma Can Predict Recurrence and Survival in Patients with Urothelial Carcinoma of the Bladder. *Cancers* 13 (10), 2377. doi:10.3390/cancers13102377

Dong, B., Liang, J., Li, D., Song, W., Zhao, S., Ma, Y., et al. (2021). Tumor Expression Profile Analysis Developed and Validated a Prognostic Model Based on Immune-Related Genes in Bladder Cancer. *Front. Genet.* 12, 696912. doi:10.3389/fgene.2021.696912

Dzobo, K. (2020). Taking a Full Snapshot of Cancer Biology: Deciphering the Tumor Microenvironment for Effective Cancer Therapy in the Oncology Clinic. *OMICS: A J. Integr. Biol.* 24 (4), 175–179. doi:10.1089/omi.2020.0019

Estrada, K., Froelich, S., Wuster, A., Bauer, C. R., Sterling, T., Clark, W. T., et al. (2021). Identifying Therapeutic Drug Targets Using Bidirectional Effect Genes. *Nat. Commun.* 12 (1), 2224. doi:10.1038/s41467-021-21843-8

Global Burden of Disease Cancer CollaborationFitzmaurice, C., Abate, D., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdel-Rahman, O., et al. (2019). Global,

Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol.* 5 (12), 1749–1768. doi:10.1001/jamaoncol.2019.2996

Gao, L., Wang, Q., Ren, W., Zheng, J., Li, S., Dou, Z., et al. (2020). The RBP1-CKAP4 axis Activates Oncogenic Autophagy and Promotes Cancer Progression in Oral Squamous Cell Carcinoma. *Cell Death Dis* 11 (6), 488. doi:10.1038/s41419-020-2693-8

Gardner, W., Cutts, S. M., Phillips, D. R., and Pigram, P. J. (2021). Understanding Mass Spectrometry Images: Complexity to Clarity with Machine Learning. *Biopolymers* 112 (4), e23400. doi:10.1002/bip.23400

Guan, X., Xu, Z.-Y., Chen, R., Qin, J.-J., and Cheng, X.-D. (2021). Identification of an Immune Gene-Associated Prognostic Signature and its Association with a Poor Prognosis in Gastric Cancer Patients. *Front. Oncol.* 10, 629909. doi:10.3389/fonc.2020.629909

Kumari, S., Advani, D., Sharma, S., Ambasta, R. K., and Kumar, P. (2021). Combinatorial Therapy in Tumor Microenvironment: Where Do We Stand? *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1876 (2), 188585. doi:10.1016/j.bbcan.2021.188585

Liu, J., Ma, H., Meng, L., Liu, X., Lv, Z., Zhang, Y., et al. (2021). Construction and External Validation of a Ferroptosis-Related Gene Signature of Predictive Value for the Overall Survival in Bladder Cancer. *Front. Mol. Biosci.* 8, 675651. doi:10.3389/fmolb.2021.675651

Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., et al. (2017). Cistrome Data Browser: a Data portal for ChIP-Seq and Chromatin Accessibility Data in Human and Mouse. *Nucleic Acids Res.* 45 (D1), D658–D662. doi:10.1093/nar/gkw983

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Peters, A. L., Stunz, L. L., and Bishop, G. A. (2009). CD40 and Autoimmunity: the Dark Side of a Great Activator. *Semin. Immunol.* 21 (5), 293–300. doi:10.1016/j.smim.2009.05.012

Petzold, A. P., Lubianca, F. N., Passos, L. G., Keppler, C. K., Becker, N. B., Viera, C. d. M., et al. (2021). The Impact of Preoperative Immune Checkpoint Inhibitors on Kidney and Bladder Cancer Surgeries: a Systematic Review☆. *Curr. Probl. Cancer*, 100765. doi:10.1016/j.currproblcancer.2021.100765

Qu, G., Liu, Z., Yang, G., Xu, Y., Xiang, M., and Tang, C. (2021). Development of a Prognostic index and Screening of Prognosis Related Genes Based on an Immunogenomic Landscape Analysis of Bladder Cancer. *Aging* 13 (8), 12099–12112. doi:10.18632/aging.202917

Ready, N., Hellmann, M. D., Awad, M. M., Otterson, G. A., Gutierrez, M., Gainor, J. F., et al. (2019). First-Line Nivolumab Plus Ipilimumab in Advanced Non-small-cell Lung Cancer (CheckMate 568): Outcomes by Programmed Death Ligand 1 and Tumor Mutational Burden as Biomarkers. *Jco* 37 (12), 992–1000. doi:10.1200/JCO.18.01042

Ren, J., Du, Y., Li, S., Ma, S., Jiang, Y., and Wu, C. (2019). Robust Network-Based Regularization and Variable Selection for High-Dimensional Genomic Data in Cancer Prognosis. *Genet. Epidemiol.* 43 (3), 276–291. doi:10.1002/gepi.22194

Ricci, R., Martini, M., Cenci, T., Carbone, A., Lanza, P., Biondi, A., et al. (2015). PDGFRA-mutant Syndrome. *Mod. Pathol.* 28 (7), 954–964. doi:10.1038/modpathol.2015.56

Richters, A., Aben, K. K. H., and Kiemeney, L. A. L. M. (2020). The Global burden of Urinary Bladder Cancer: an Update. *World J. Urol.* 38 (8), 1895–1904. doi:10.1007/s00345-019-02984-4

Seidel, M. G., Boztug, K., and Haas, O. A. (2016). Immune Dysregulation Syndromes (IPEX, CD27 Deficiency, and Others): Always Doomed from the Start? *J. Clin. Immunol.* 36 (1), 6–7. doi:10.1007/s10875-015-0218-5

Spenlé, C., Loustau, T., Burckel, H., Riegel, G., Abou Faycal, C., Li, C., et al. (2021). Impact of Tenascin-C on Radiotherapy in a Novel Syngeneic Oral Squamous Cell Carcinoma Model with Spontaneous Dissemination to the Lymph Nodes. *Front. Immunol.* 12, 636108. doi:10.3389/fimmu.2021.636108

Sylvester, R. J., van der Meijden, A. P. M., Oosterlinck, W., Witjes, J. A., Bouffioux, C., Denis, L., et al. (2006). Predicting Recurrence and Progression in Individual Patients with Stage Ta T1 Bladder Cancer Using EORTC Risk Tables: a Combined Analysis of 2596 Patients from Seven EORTC Trials. *Eur. Urol.* 49 (3), 466–477. doi:10.1016/j.eururo.2005.12.031

Tosev, G., Wahafu, W., Reimold, P., Damgov, I., Schwab, C., Aksoy, C., et al. (2021). Detection of PD-L1 in the Urine of Patients with Urothelial Carcinoma of the Bladder. *Sci. Rep.* 11 (1), 14244. doi:10.1038/s41598-021-93754-z

Tran, L., Xiao, J.-F., Agarwal, N., Duex, J. E., and Theodorescu, D. (2021). Advances in Bladder Cancer Biology and Therapy. *Nat. Rev. Cancer* 21 (2), 104–121. doi:10.1038/s41568-020-00313-1

Vackova, J., Polakova, I., Johari, S. D., and Smahel, M. (2021). CD80 Expression on Tumor Cells Alters Tumor Microenvironment and Efficacy of Cancer Immunotherapy by CTLA-4 Blockade. *Cancers* 13 (8), 1935. doi:10.3390/cancers13081935

van Rhijn, B. W. G., Burger, M., Lotan, Y., Solsona, E., Stief, C. G., Sylvester, R. J., et al. (2009). Recurrence and Progression of Disease in Non-muscle-invasive Bladder Cancer: from Epidemiology to Treatment Strategy. *Eur. Urol.* 56 (3), 430–442. doi:10.1016/j.eururo.2009.06.028

Wang, Z., Tu, L., Chen, M., and Tong, S. (2021). Identification of a Tumor Microenvironment-Related Seven-Gene Signature for Predicting Prognosis in Bladder Cancer. *BMC Cancer* 21 (1), 692. doi:10.1186/s12885-021-08447-7

Witjes, J. A., Bruins, H. M., Cathomas, R., Compérat, E. M., Cowan, N. C., Gakis, G., et al. (2021). European Association of Urology Guidelines on Muscle-Invasive and Metastatic Bladder Cancer: Summary of the 2020 Guidelines. *Eur. Urol.* 79 (1), 82–104. doi:10.1016/j.eururo.2020.03.055

Wu, C., and Ma, S. (2015). A Selective Review of Robust Variable Selection with Applications in Bioinformatics. *Brief Bioinform* 16 (5), 873–883. doi:10.1093/bib/bbu046

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Yu, M., Zhao, H., Miao, Y., Luo, S.-Z., and Xue, S. (2021). Virtual Evolution of HVEM Segment for Checkpoint Inhibitor Discovery. *Ijms* 22 (12), 6638. doi:10.3390/ijms22126638

Zalcberg, J. R. (2021). Ripretinib for the Treatment of Advanced Gastrointestinal Stromal Tumor. *Therap Adv. Gastroenterol.* 14, 175628482110081. doi:10.1177/17562848211008177

Zhang, L.-H., Li, L.-Q., Zhan, Y.-H., Zhu, Z.-W., and Zhang, X.-P. (2021). Identification of an IRGP Signature to Predict Prognosis and Immunotherapeutic Efficiency in Bladder Cancer. *Front. Mol. Biosci.* 8, 607090. doi:10.3389/fmolb.2021.607090

Zhu, Y. D., and Lu, M. Y. (2018). Increased Expression of TNFRSF14 Indicates Good Prognosis and Inhibits Bladder Cancer Proliferation by Promoting Apoptosis. *Mol. Med. Rep.* 18 (3), 3403–3410. doi:10.3892/mmr.2018.9306

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 December 2021
doi: 10.3389/fgene.2021.791778

The Causal Effects of Primary Biliary Cholangitis on Thyroid Dysfunction: A Two-Sample Mendelian Randomization Study

Peng Huang[1†], *Yuqing Hou*[1†], *Yixin Zou*[1], *Xiangyu Ye*[1], *Rongbin Yu*[1] *and Sheng Yang*[2*]

[1]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China, [2]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China

Background: Primary biliary cholangitis (PBC) is an autoimmune disease and is often accompanied by thyroid dysfunction. Understanding the potential causal relationship between PBC and thyroid dysfunction is helpful to explore the pathogenesis of PBC and to develop strategies for the prevention and treatment of PBC and its complications.

Methods: We used a two-sample Mendelian randomization (MR) method to estimate the potential causal effect of PBC on the risk of autoimmune thyroid disease (AITD), thyroid-stimulating hormone (TSH) and free thyroxine (FT4), hyperthyroidism, hypothyroidism, and thyroid cancer (TC) in the European population. We collected seven datasets of PBC and related traits to perform a series MR analysis and performed extensive sensitivity analyses to ensure the reliability of our results.

Results: Using a sensitivity analysis, we found that PBC was a risk factor for AITD, TSH, hypothyroidism, and TC with odds ratio (OR) of 1.002 (95% CI: 1.000–1.005, $p = 0.042$), 1.016 (95% CI: 1.006–1.027, $p = 0.002$), 1.068 (95% CI: 1.022–1.115, $p = 0.003$), and 1.106 (95% CI: 1.019–1.120, $p = 0.042$), respectively. Interestingly, using reverse-direction MR analysis, we also found that AITD had a significant potential causal association with PBC with an OR of 0.021 ($p = 5.10E{-}4$) and that the other two had no significant causal relation on PBC.

Conclusion: PBC causes thyroid dysfunction, specifically as AITD, mild hypothyroidism, and TC. The potential causal relationship between PBC and thyroid dysfunction provides a new direction for the etiology of PBC.

Keywords: thyroid dysfunction, hypothyroidism, thyroid cancer, two-sample Mendelian randomization, genome-wide association study

OPEN ACCESS

Edited by:
Jiajie Peng,
Northwestern Polytechnical University, China

Reviewed by:
Mingwang Shen,
Xi'an Jiaotong University, China
Xiaomei Ma,
Independent Researcher, Zhengzhou, China

***Correspondence:**
Sheng Yang
yangsheng@njmu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *09 October 2021*
Accepted: *15 November 2021*
Published: *10 December 2021*

Citation:
Huang P, Hou Y, Zou Y, Ye X, Yu R and Yang S (2021) The Causal Effects of Primary Biliary Cholangitis on Thyroid Dysfunction: A Two-Sample Mendelian Randomization Study. Front. Genet. 12:791778. doi: 10.3389/fgene.2021.791778

INTRODUCTION

Primary biliary cholangitis (PBC) is an autoimmune cholestatic liver disease with a progressive disease (Lleo et al., 2017). Its prevalence and annual incidence rate are from 6.7 to 492 cases and from 0.7 to 49 cases per million inhabitants, respectively (Delgado et al., 2012). In addition, some population-based studies that investigate the incidence and prevalence of PBC are increasing year by year (Carey et al., 2015; Rosa et al., 2018; Lindor et al., 2019). Even worse, similar to other

autoimmune diseases, the pathogenesis of PBC is complex and multifactorial, which results in no effective treatment for PBC. PBC can lead to cirrhosis, liver cancer, liver failure, and death within 10 years (Pratt, 2016; Younossi et al., 2019). With increasing prevalence and serious complications, it is worthy of our further study to explore the possible pathogenesis of PBC.

As known, up to 73% of PBC patients have extrahepatic manifestations (i.e., Sjogren's syndrome, thyroid dysfunction, and systemic sclerosis) whose high incidence declines the quality of life (Chalifoux et al., 2017). Among them, thyroid dysfunction occurs in 5.6%–23.6% of PBC patients, which is obviously larger than that in the individuals without PBC (Huang and Liaw, 1995; Gershwin et al., 2005; Silveira et al., 2009; Chalifoux et al., 2017). Thyroid dysfunction results from excessive or insufficient production of thyroid hormone regulating human growth, neuron development, reproduction, and energy metabolism and may lead to various thyroid diseases such as hypothyroidism (Taylor et al., 2018). Studies have shown that some patients have symptoms of hypothyroidism and PBC symptoms simultaneously, and the incidence of hypothyroidism increases in patients with PBC (Elta et al., 1983). It is reported that levels of serum thyroid-stimulating hormone (TSH) and average serum free thyroxine (FT4) are higher in PBC patients (Schussler et al., 1978). The two kinds of hormone imbalance also occur in cirrhotic patients (Vincken et al., 2017; Punekar et al., 2018). In addition, patients with PBC, an autoimmune liver disease (AILD), are at higher risk for other autoimmune diseases, including autoimmune thyroid disease (AITD), Hashimoto's thyroiditis, and Graves's thyroiditis (Floreani et al., 2015; Suzuki et al., 2016; Zeng et al., 2020). However, the abovementioned relationships between PBC and thyroid dysfunction are obtained based on observational studies, in which reverse causality, selection bias, and especially unobserved confounding factors might mask true causal relationships. It is essential to further investigate the causal association underlying these correlations.

Mendelian randomization (MR) is widely used for causal inference in observational studies by treating single-nucleotide polymorphisms (SNPs) as instrumental variables (IVs) (Emdin et al., 2017). According to Mendel's law of inheritance, alleles are transmitted randomly from parents to offspring during meiosis without interference from external factors (Reyna and Pickler, 1999). Therefore, MR has a natural advantage in determining causal relationships by removing the unobserved confounding. In addition, two-sample MR, requiring the exposure and outcome measured in independent but homogeneous samples, is accessible for the abundant resources and availability of summary statistics of genome-wide association studies (GWASs) (Evans and Davey Smith, 2015; Watanabe et al., 2019). MR is established on the basis that if a causal relationship exists between PBC and thyroid dysfunction, the SNP related to PBC will also be related to thyroid dysfunction through the occurrence of PBC, in which case the IVs only associate with PBC, and MR can help establish a causal relationship between PBC and thyroid function (Davey Smith and Hemani, 2014). Also, two-sample MR has been used to explore the causal relationship between thyroid function and breast cancer (Yuan et al., 2020), atrial fibrillation (Ellervik et al., 2019), and blood lipid profile (Wang et al., 2021).

Here, we comprehensively investigate the potential causal relationship between PBC and thyroid function. Specifically, we use seven large-scale GWAS summary statistics in the European population on PBC and thyroid indicators and disorders, including AITD, TSH, FT4, hyperthyroidism, hypothyroidism, and thyroid cancer (TC), to perform a series of two-sample MR. Furthermore, we also perform several sensitivity analyses, including the heterogeneity test, pleiotropy test, leave-one-out (LOO) test, and reverse-direction MR analyses to ensure the reliability of our results.

METHODS

Data Collection

We collected seven datasets on PBC and related traits, including one PBC dataset (Cordell et al., 2015), one AITD dataset (including both Hashimoto's thyroiditis and Graves' disease) (Glanville et al., 2021), four datasets (for TSH, FT4, hyperthyroidism, and hypothyroidism) from The ThyroidOmics Consortium (Teumer et al., 2018), and one TC dataset (Rashkin et al., 2020). Specifically, PBC dataset contained 13,239 individuals (Prev. = 0.209); AITD dataset contained 324,933 individuals (Prev. = 0.003); TSH dataset contained 54,288 individuals; FT4 dataset contained 49,269 individuals; hyperthyroidism dataset included 51,668 individuals (Prev. = 0.626); hypothyroidism dataset included 53,241 individuals (Prev. = 0.036); and TC dataset included 411,112 individuals (Prev. = 0.002). All summary data came from the European population. Then, we filter out SNPs 1) with INFO < 0.6, 2) with minor allele fraction (MAF) <0.01, 3) with palindrome alleles, and 4) whose odds ratio (OR) was larger or smaller than the mean ± 3 SD. Finally, we obtained 1,134,141, 9,390,112, 7,666,442, 7,138,715, 7,138,916, 7,191,562, and 9,291,956 SNPs for the seven traits. In addition, we used linkage disequilibrium score regression (LDSC) (v1.0.1) to estimate heritability (h^2) for each dataset. We set the population prevalence (--pop-prev) for the five diseases (PBC, AITD, hyperthyroidism, hypothyroidism, and TC) to 0.209, 0.003, 0.036, 0.063, and 0.002 to estimate liability heritability. The detailed information of the seven datasets is shown in **Supplementary Table S1**.

Instrumental Variable Selections

A crucial step of MR was to choose appropriate genetic variants to serve as valid IVs for PBC. Based on the above datasets, we followed the strict screening procedures in other previous MR studies to select IVs (Zeng et al., 2019; Dong et al., 2021) (**Figure 1**). First, we retained 773 variants for PBC with a *p*-value smaller than 5.00E−8. Second, we removed 748 highly correlated variants with r^2 greater than 0.001 in the range of 10 Mb. In addition, we ensured that each alternative SNP selected as IV was strongly associated with PBC. According to the previous research (Zeng et al., 2019), we calculated the *F* statistic to find weak IVs, and no variant was excluded with a minimum *F* statistic of 30.16. Finally, we only kept a total of 25

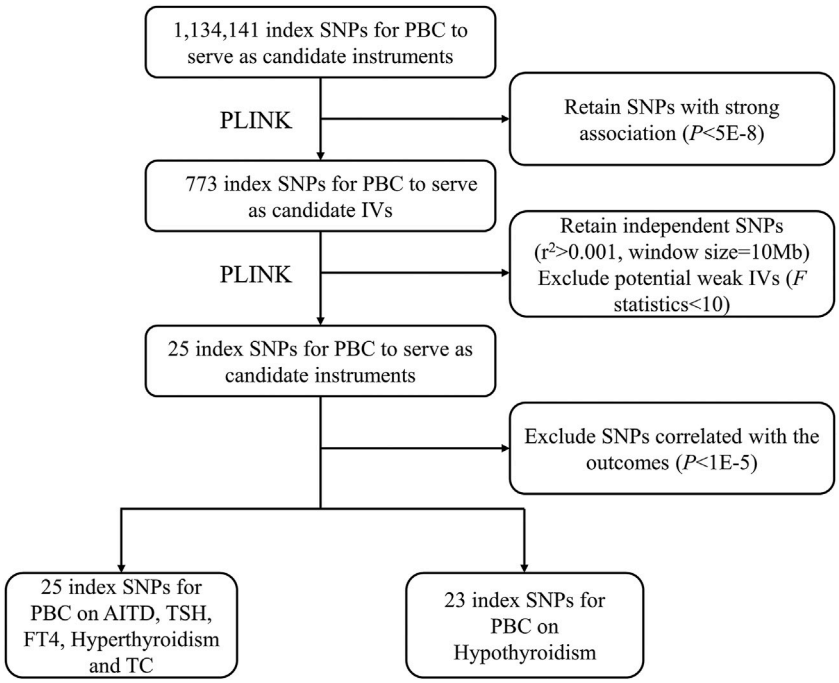

FIGURE 1 | The flowchart for IV selection. The flowchart shows the selection process of PBC IVs to estimate the causal effects on AITD, TSH, FT4, hyperthyroidism, hypothyroidism, and TC. First, we use $p < 5.00E{-}8$ to select index SNPs to ensure that they strongly associate with PBC. Second, we use $r^2 > 0.001$ in the range of 10 Mb to select independent index SNPs. We treat the EUR of 1000 Genomes Project as the reference panel. The first two steps are completed by PLINK. Finally, we obtain 25 IVs on AITD, TSH, FT4, hyperthyroidism, and TC and 23 IVs on hypothyroidism. IV, instrumental variable; PBC, primary biliary cholangitis; AITD, autoimmune thyroid disease; TSH, thyroid-stimulating hormone; FT4, free thyroxine; TC, thyroid cancer; SNP, single-nucleotide polymorphism.

independent candidate IVs to study the causal relationship between PBC and the other six traits. The details of these IVs are shown in **Supplementary Table S2**.

We carried out three two-sample MR analyses, including fixed-effects and random-effects inverse variance weighting (IVW), MR-Egger, and weighted median (WM) methods, to estimate the potential causal effect of PBC on the six traits (Bowden et al., 2015; Bowden et al., 2016a). Without consideration for the intercept term, IVW regarded the reciprocal of the outcome variance (the square of SE) as the weight. Under the assumption of IVW, we consider that IVs are not pleiotropic. Therefore, we must ensure that these IVs are not pleiotropic when using the IVW method; otherwise, the results were biased (Bowden et al., 2015). Different from IVW, MR-Egger used an intercept term to measure the horizontal pleiotropy between these IVs (Bowden et al., 2016b). The WM method assumed that variables that account for at least 50% of the total IVs were valid, so the causal effects can be estimated consistently (Bowden et al., 2016a).

Sensitivity Analysis

Following methods in previous studies (Noyce et al., 2017; Zeng and Zhou, 2019), we performed a sensitivity analysis to evaluate the potential violations of the model assumptions in the MR analysis: 1) heterogeneity test, 2) pleiotropic test, and 3) LOO test. First, heterogeneity analysis estimates the heterogeneity between IVs. If the heterogeneity exists, it would be hard to combine the IVs directly. Second, if IVs can directly affect the results without exposure factors, then they violate the idea of MR; that is, the level of pleiotropy in the test results will lead to serious deviations in MR (Hemani et al., 2018; Ong and MacGregor, 2019). We use MR pleiotropy residual sum and outlier (MR-PRESSO) to find outliers and test the level of pleiotropy. For more verification, we still use the MR-Egger intercept to test the pleiotropy. Finally, the LOO test refers to gradually removing each SNP, calculating the meta effect of the remaining SNPs, and observing whether the result significantly changed after removing each SNP. Ideally, no significant difference meant a robust result (Noyce et al., 2017). All the analyses are performed by R software (v4.1.1). Specially, we used TwoSampleMR R package (v0.5.6) to perform MR analysis. The statistical significance level was set to 0.05 throughout our study.

Reverse-Direction Mendelian Randomization Analyses

We also performed reverse-direction MR to assess potential reverse causal effects of AITD, TSH, FT4, hyperthyroidism, hypothyroidism, and TC on PBC. Following methods in previous literature (Savage et al., 2018; Dong et al., 2021), for each exposure, we used the clumping algorithm in PLINK (Chang et al., 2015) to select independent SNPs for each trait (r^2 threshold = 0.001, window size = 10 Mb, and $p < 5.00E{-}8$). Finally, we obtained two IVs for AITD, 38 IVs for TSH, 17 IVs for FT4, 7 IVs for hyperthyroidism, 6

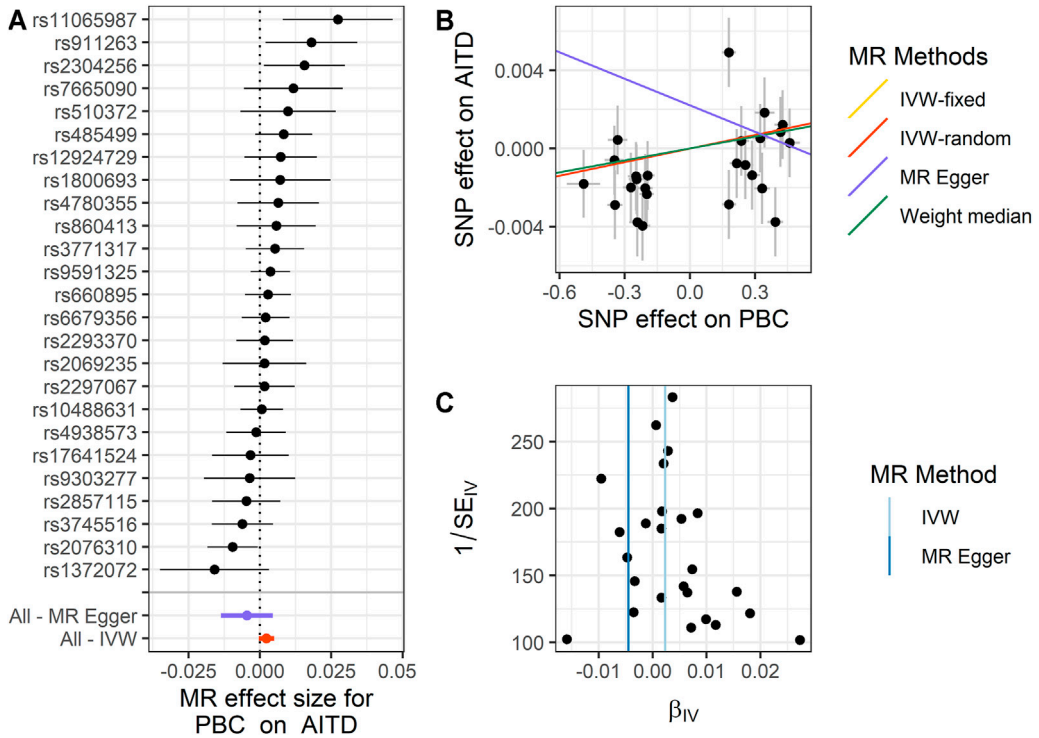

FIGURE 2 | Summary of the MR analysis for PBC on AITD. **(A)** MR effect size of each IV, MR-Egger, and IVW. **(B)** The scatter plot of causal effects of PBC on AITD. We use vertical and horizontal black lines to show 95% CI of the estimated effect of IVs on PBC (x-axis) and that on AITD (y-axis), respectively. We use the red line to show the IVW random-effects model. **(C)** The funnel plot of the causal effect of PBC on AITD. Each point represents the estimated causal effect of each IV. The vertical dark blue line represents the causal effect estimate obtained using the MR-Egger method; the light blue line represents the causal effect estimate obtained using the IVW method. MR, Mendelian randomization; PBC, primary biliary cholangitis; AITD, autoimmune thyroid disease; IV, instrumental variable; IVW, inverse variance weighting.

IVs for hypothyroidism, and 3 IVs for TC. We used these IVs of six traits to perform reverse causal inferences on PBC to assess potential reverse causal effects. The reverse-direction MR analysis process was the same as previously described.

RESULTS

Summary of Genome-Wide Association Study Data

We estimated the heritability for each trait. Specifically, the genetic inflation factor (λ_{gc}) of PBC is 1.050 (LDSC intercept: 1.003); λ_{gc} of AITD is 0.999 (LDSC intercept: 0.999); λ_{gc} of TSH is 1.077 (LDSC intercept: 1.035); λ_{gc} of FT4 is 1.111 (LDSC intercept: 1.014); λ_{gc} of hyperthyroidism is 1.029 (LDSC intercept: 1.113); λ_{gc} of hypothyroidism is 1.044 (LDSC intercept: 1.083); and λ_{gc} of TC is 1.008 (LDSC intercept: 0.999). With the use of GWAS summary statistics and 1000 Genomes Project (1000 GP) EUR reference panel, the SNP-based liability heritability for PBC, AITD, and TC is 595.942, 0.012, and 0.103, respectively. The observed heritability for PBC, AITD, TSH, FT4, and TC is 0.378, 0.003, 0.125, 0.152, and 0.002, respectively (**Supplementary Table S1**). We used the Manhattan plot to show the GWAS results for seven traits (**Supplementary Figure S1**).

Mendelian Randomization Analysis

We performed MR analysis on the IVs of PBC selected on six traits. Except for hypothyroidism, which only had 23 IVs, the other five traits were all 25 IVs. Based on different assumptions, we estimated the potential causal effects of four models, including IVW (fixed- and random-effects models), MR-Egger, and WM. And we use forest plots to show the causal relationship of a single IV in each trait, scatter plots to show the overall fitting causal effects between PBC and the traits, and funnel plots to show the relationship between the effect of the MR model and the effect of each SNP (**Figures 2–4**; **Supplementary Figures S2–S4**, **Supplementary Tables S3–S8**). For the causal effect for the six traits, we should use the result of the sensitivity analysis to determine whether the analysis result is significant.

For PBC on AITD, we observe a positive causal effect. The estimated OR from fixed-effects IVW method is 1.002 (95% CI: 1.000–1.005, p = 0.042). However, the result of the random-effects IVW method (OR = 1.002, 95% CI: 0.999–1.005, p = 0.092) is similar to that of fixed-effects IVW, but it is not significant. The result of WM (OR = 1.002, 95% CI: 1.000–1.005, p = 0.196) and MR-Egger (OR = 0.995, 95% CI: 0.987–1.005, p = 0.339) is similar to that of the random-effects IVW method. The above results indicate that AITD would increase with the increase of PBC risk. The details are shown in **Figure 2** and **Supplementary Table S3**.

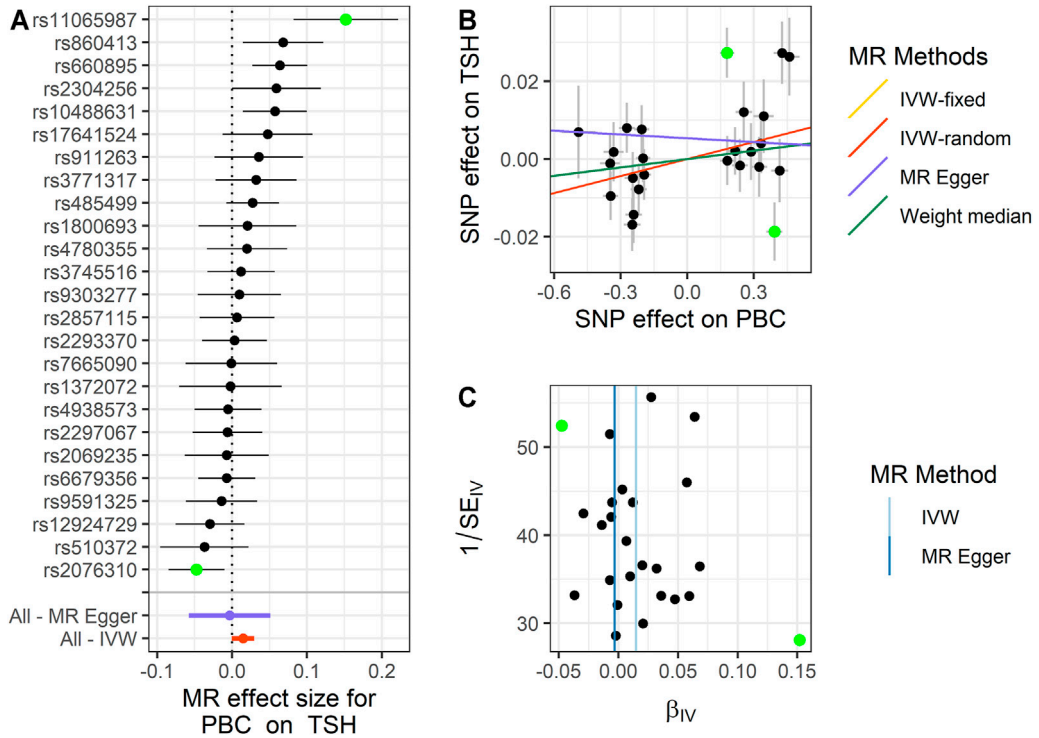

FIGURE 3 | Summary of the MR analysis for PBC on TSH. **(A)** MR effect size of each IV, MR-Egger, and IVW. **(B)** The scatter plot of causal effects of PBC on TSH. We use vertical and horizontal black lines to show 95% CI of the estimated effect of IVs on PBC (x-axis) and that on TSH (y-axis), respectively. We use the red line to show the IVW random-effects model. The potential SNP outlier (rs11065987 and rs2076310) is highlighted in green. **(C)** The funnel plot of the causal effect of PBC on TSH. Each point represents the estimated causal effect of each IV. The vertical dark blue line represents the causal effect estimate obtained using the MR-Egger method; the light blue line represents the causal effect estimate obtained using the IVW method. The potential SNP outlier (rs11065987 and rs2076310) is highlighted in green. MR, Mendelian randomization; PBC, primary biliary cholangitis; TSH, thyroid-stimulating hormone; IV, instrumental variable; IVW, inverse variance weighting; SNP, single-nucleotide polymorphism.

For PBC on TSH, we also observe a positive causal effect. The estimated OR from fixed-effects IVW method is 1.015 (95% CI: 1.005–1.025, p = 0.003). The result of the random-effects IVW method (OR = 1.015, 95% CI: 1.000–1.030, p = 0.056) is similar to that of fixed-effects IVW, but it is not significant. The result of WM (OR = 1.007, 95% CI: 0.991–1.024, p = 0.362) and MR-Egger (OR = 0.997, 95% CI: 0.944–1.052, p = 0.908) is similar to that of the random-effects IVW method. The above results indicate that TSH would increase with the increase of PBC risk. The details are shown in **Figure 3** and **Supplementary Table S4**.

For PBC on FT4, we failed to define a significant causal effect. The estimated OR from fixed-effects IVW method is 1.005 (95% CI: 0.994–1.015, p = 0.375). And the result of the random-effects IVW method (OR = 1.005, 95% CI: 0.994–1.015, p = 0.380) is similar to that of fixed-effects IVW. The result of WM (OR = 1.007, 95% CI: 0.992–1.022, p = 0.372) and MR-Egger (OR = 0.997, 95% CI: 0.960–1.036, p = 0.893) is similar to the above conclusion. The above results indicate that FT4 would increase with the increase of PBC risk, but none of them is significant. The details are shown in **Supplementary Figure S2** and **Supplementary Table S5**.

For PBC on hyperthyroidism, we failed to define any significant causal effect using the four models. The estimated OR from fixed-effects IVW method is 0.984 (95% CI: 0.937–1.034, p = 0.534). And the result of the random-effects IVW method (OR = 0.984, 95% CI: 0.938–1.034, p = 0.524) is not significant, which is similar to that of fixed-effects IVW. The results of WM (OR = 1.007, 95% CI: 0.939–1.081, p = 0.836) and MR-Egger (OR = 0.933, 95% CI: 0.779–1.117, p = 0.457) are both not significant. The details are shown in **Supplementary Figure S3** and **Supplementary Table S6**.

For PBC on hypothyroidism, in these 23 IVs, we observed the positive causal effect of PBC on hypothyroidism. Note that we only define the significant result from fixed-effects IVW method (OR = 1.044, 95% CI: 1.001–1.089, p = 0.044), rather than the result of the random-effects IVW method (OR = 1.044, 95% CI: 0.989–1.103, p = 0.122), the result of WM (OR = 1.051, 95% CI: 0.987–1.119, p = 0.116), and the result of MR-Egger (OR = 0.912, 95% CI: 0.744–1.118, p = 0.385). We should use the result of the sensitivity analysis to check for the outliers and determine whether the analysis result is representative. The details are shown in **Figure 4** and **Supplementary Table S7**.

For PBC on TC, the result of PBC on TC is similar to that of hypothyroidism. The only significant result was from fixed-effects IVW method (OR = 1.106, 95% CI: 1.019–1.120, p = 0.042) rather than the result of the random-effects IVW method

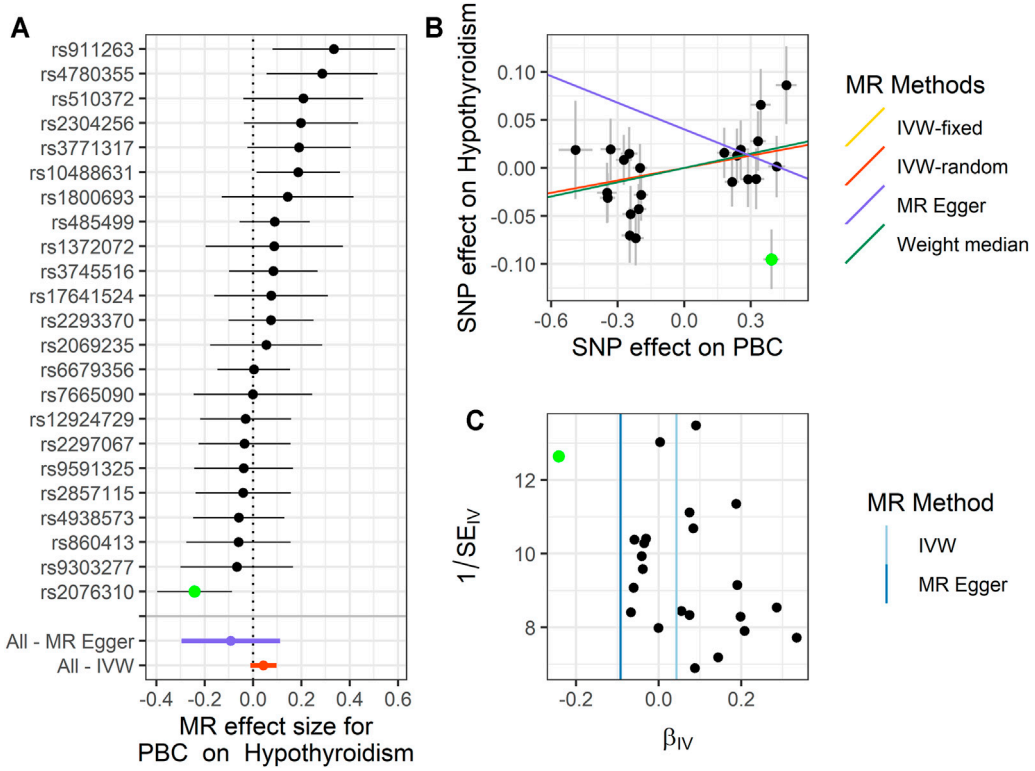

FIGURE 4 | Summary of the MR analysis for PBC on hypothyroidism. **(A)** MR effect size of each IV, MR-Egger, and IVW. **(B)** The scatter plot of causal effects of PBC on hypothyroidism. We use vertical and horizontal black lines to show 95% CI of the estimated effect of IVs on PBC (x-axis) and that on hypothyroidism (y-axis), respectively. We use the red line to show the IVW random-effects model. The potential SNP outliers (rs2076310) are highlighted in green. **(C)** The funnel plot of the causal effect of PBC on hypothyroidism. Each point represents the estimated causal effect of each IV. The vertical dark blue line represents the causal effect estimate obtained using the MR-Egger method; the light blue line represents the causal effect estimate obtained using the IVW method. The potential outliers (rs2076310) are highlighted in green. MR, Mendelian randomization; PBC, primary biliary cholangitis; IV, instrumental variable; IVW, inverse variance weighting; SNP, single-nucleotide polymorphism.

(OR = 1.106, 95% CI: 0.990–1.235, $p = 0.074$), the result of WM (OR = 1.137, 95% CI: 0.998–1.295, $p = 0.054$), and the result of MR-Egger (OR = 1.243, 95% CI: 0.836–1.850, $p = 0.294$). The details are shown in **Supplementary Figure S4** and **Supplementary Table S8**.

Sensitivity Analyses

We performed extensive sensitivity analyses to validate the results of the MR analysis, which are mainly of heterogeneity analysis and pleiotropic analysis on these IVs. We performed LOO analysis only for the significant causal effect. Our purpose was to explore whether the results obtained were robust, whether there was potential bias (such as pleiotropy and data heterogeneity), and whether there was a certain IV that seriously affects the outcome variable.

First, we conducted a heterogeneity analysis. Based on IVW, we found that TSH, hypothyroidism, and TC were heterogeneous as compared with AITD, FT4, and hypothyroidism. The P_Q of TSH, hyperthyroidism, and TC was 1.71E−4, 0.021, and 0.007, respectively. The P_Q of the remaining three traits was larger than 0.05. In order to reduce the heterogeneity, we chose to perform MR-PRESSO analysis to find and eliminate the outliers and can also test the pleiotropy of IVs.

Next, we performed MR-PRESSO analysis and LOO test to ensure the validation for MR analysis. For TSH, we found rs11065987 (beta = 0.027, $p = 2.30E{-}5$) and rs2076310 (beta = −0.019, $p = 0.013$) might be outliers that have affected the causal effect of IVs. After they were removed, the p-value of MR-PRESSO Global test changed from 0.001 to 0.098, which indicated that the pleiotropy was eliminated, and the P_Q of TSH was also changed to 0.089, indicating that the heterogeneity has been eliminated. The result of the LOO test was significant for rs2076310 ($P_{LOO} = 0.008$) and not significant for rs11065987 ($P_{LOO} = 0.078$). Therefore, we choose the result of the fixed-effects IVW method after removing the outliers as the significant causal effect for PBC on TSH. For hypothyroidism, we define that rs2076310 (beta = −0.095, $p = 0.002$) might be an outlier that has affected the causal effect of IVs. After it was removed, the p-value of MR-PRESSO Global test changed from 0.022 to 0.345, which indicated that the pleiotropy has been eliminated, and the P_Q of hypothyroidism was also changed to 0.321, indicating that the heterogeneity has been eliminated. The result of the LOO test for this outlier is the same as the above

result (P_{LOO} = 0.005). Therefore, we chose the result of the fixed-effects IVW method after removing the outlier as the significant causal effect of PBC on hypothyroidism. And for TC, we did not identify any outliers.

Finally, we used the MR-Egger intercept to estimate pleiotropy. We defined no significant pleiotropy in six potential causal relationships. After the outliers were removed, the *p*-value of the MR-Egger intercept increased.

To sum up, we used the result from the fixed-effects IVW method to represent the causal effect of PBC on AITD (OR = 1.002, 95% CI: 1.000–1.005, p = 0.042); the result from fixed-effects IVW method with outlier excluded was used to represent the causal effect of PBC on TSH (OR = 1.016, 95% CI: 1.006–1.027, p = 0.002); and the result from fixed-effects IVW method with outlier excluded was used to represent the causal effect of PBC on hypothyroidism (OR = 1.068, 95% CI: 1.022–1.115, p = 0.003). The remaining three traits were not significant.

Reverse-Direction Mendelian Randomization Analysis

In order to identify potential confounding factors that mislead the direction of causal effects, we performed reverse-direction MR (**Supplementary Figures S5–10**). We found that AITD and TC have a significant potential causal association with PBC with the random-effects IVW method, while the causal effects for TSH, FT4, hyperthyroidism, and hypothyroidism on PBC are not significant. Specifically, using the random-effects IVW method, the estimated OR for AITD and TC on PBC is 0.021 (p = 5.10E−4) and 1.026 (p = 0.011), respectively (**Supplementary Tables S3, S8**). Note that the results might be not reliable for the small number of IVs (Dong et al., 2021).

DISCUSSION

Here, we performed a comprehensive two-sample MR analysis to illustrate the potential causality between PBC and thyroid dysfunction. After a series of sensitivity analyses, we found that PBC significantly results in the occurrence of AITD (OR=1.002) and hypothyroidism (OR = 1.068) and that PBC significantly causes the increase of TSH level (OR = 1.016). Our findings provided an exploration direction for the occurrence of thyroid dysfunction in PBC patients, contributed to the treatment of thyroid diseases in PBC patients, and improved the quality of life for PBC patients. As expected, our results are consistent with previous observational population-based studies. For the potential causal relation of hypothyroidism, emerging evidences indicate that PBC is often with the occurrence of AITD (Crowe et al., 1980; Floreani et al., 2015; Patil et al., 2021) and hypothyroidism (Crowe et al., 1980; Elta et al., 1983) and the increase of TSH, one of the main signs of hypothyroidism (Patil et al., 2021).

For PBC on AITD, we define that PBC and AITD might be mutual cause-and-effect factors in both MR and reverse-direction MR analyses. Consistent with our findings, emerging epidemiological studies have shown that genetic components are important in the pathogenesis of Hashimoto's thyroiditis (Paknys et al., 2009). The occurrence of PBC and AITD might be caused by environmental and genetic factors, such as intestinal flora (Fenneman et al., 2020), estrogen (Qin et al., 2018), gene-mediated immunodeficiency, and synergy between each other (Milette et al., 2019).

For PBC on hypothyroidism, Garber et al. also showed that PBC causes mild hypothyroidism, manifesting as only increasing TSH and normal FT4 levels (Garber et al., 2012). This finding is consistent with our results, that is, significant causal relation of PBC on TSH and insignificant causal relation of PBC on FT4.

There are several assumptions for the causality for PBC on hypothyroidism. One is the interaction between thyroid hormones and the liver (Salata et al., 1985). Liver damage caused by PBC can lead to changes in the expression of the enzyme D3 that controls the activity of thyroid hormones (Gilgenkrantz and Collin de l'Hortet, 2018), which can lead to a decrease in the accumulation of active thyroid hormones (Elbers et al., 2016), can trigger hypothalamic–pituitary–thyroid regulation disorders, and can increase TSH, leading to hypothyroidism. The second is that PBC cholestasis decreases Y protein, which in turn leads to hypothyroidism (Ariza et al., 1984). Protein Y is a type of protein that is distributed in the liver and promotes the absorption of thyroid hormones by the liver; the decrease of protein Y makes the liver speed up the circulation of thyroid hormones and reduce the free thyroid hormones in the blood (Reyes et al., 1971), leading to hypothalamus–pituitary–thyroid disorders, increase in TSH, and appearance of symptoms of hypothyroidism.

For the potential causal relation for PBC on TC, few studies have reported the association between PBC and TC. We assume that PBC causes thyroid dysfunction (such as hypothyroidism and thyroiditis), which eventually progresses to TC (dos Santos Silva and Swerdlow, 1993; Pacini et al., 2012). Studies have shown that AITD is one of the risk factors for TC, and elevated TSH levels and thyroid autoimmune characteristics are defined as independent risk factors for TC (Ferrari et al., 2020). Studies have also shown that thyroid tumors mainly exhibit hypothyroidism-like symptoms and that hypothyroidism may be the basis for most TCs (Hernandez et al., 2021). Our research is consistent with previous findings and explanations.

Our research also has some limitations. First, MR analysis cannot rule out the influence of hidden and unknown confounding factors, and we cannot completely rule out the association of IVs to confounding factors. This makes the assumptions of IVs strict and demanding. Especially weak IVs should be considered in the research. Second, MR analysis only provides directions for the etiology and progress of PBC and thyroid dysfunction, which lacks the biological mechanism behind the potential causal relationship. Last, the populations of the data we analyzed are all of European descent, the final results are limited by the genes of different races, and the results may not be very applicable to Asian populations.

In conclusion, our findings show that PBC can cause thyroid dysfunction, specifically as AITD, mild hypothyroidism, and TC. The potential causal relationship between PBC and thyroid dysfunction provides a new direction for the study of the etiology and progress of PBC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: (https://www.immunobase.org/downloads/protected_data/GWAS_Data/hg19_gwas_pbc_cordell_4_20_0.tab.gz; https://transfer.sysepi.medizin.uni-greifswald.de/thyroidomics/datasets/; http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90014001-GCST90015000/GCST90014440/GCST90014440_buildGRCh37.tsv; https://github.com/Wittelab/pancancer_pleiotropy.

AUTHOR CONTRIBUTIONS

SY and YH designed the study. YZ and XY performed the dataset quality control. YZ and XY performed the data analysis. PH, YH, and YZ interpreted the analysis results. PH and YH wrote the draft manuscript. RY and SY revised the article. All authors accepted the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (no. 81703321) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

ACKNOWLEDGMENTS

We acknowledge the participants and investigators of GWAS-ALTAS for making the summary data publicly available for us.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.791778/full#supplementary-material

REFERENCES

Ariza, C. R., Frati, A. C., and Sierra, I. (1984). Hypothyroidism-Associated Cholestasis. *JAMA* 252, 2392. doi:10.1001/jama.1984.03350170010007

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44, 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016a). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40, 304–314. doi:10.1002/gepi.21965

Bowden, J., Del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016b). Assessing the Suitability of Summary Data for Two-Sample Mendelian Randomization Analyses Using MR-Egger Regression: the Role of the I2 Statistic. *Int. J. Epidemiol.* 45, dyw220–1974. doi:10.1093/ije/dyw220

Carey, E. J., Ali, A. H., and Lindor, K. D. (2015). Primary Biliary Cirrhosis. *Lancet* 386, 1565–1575. doi:10.1016/S0140-6736(15)00154-3

Chalifoux, S. L., Konyn, P. G., Choi, G., and Saab, S. (2017). Extrahepatic Manifestations of Primary Biliary Cholangitis. *Gut and Liver* 11, 771–780. doi:10.5009/gnl16365

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-Generation PLINK: Rising to the challenge of Larger and Richer Datasets. *GigaSci* 4, 7. doi:10.1186/s13742-015-0047-8

Cordell, H. J., Han, Y., Han, Y., Mells, G. F., Li, Y., Hirschfield, G. M., et al. (2015). International Genome-wide Meta-Analysis Identifies New Primary Biliary Cirrhosis Risk Loci and Targetable Pathogenic Pathways. *Nat. Commun.* 6, 8019. doi:10.1038/ncomms9019

Crowe, J. P., Christensen, E., Butler, J., Wheeler, P., Doniach, D., Keenan, J., et al. (1980). Primary Biliary Cirrhosis: the Prevalence of Hypothyroidism and its Relationship to Thyroid Autoantibodies and Sicca Syndrome. *Gastroenterology* 78, 1437–1441. doi:10.1016/s0016-5085(19)30497-4

Davey Smith, G., and Hemani, G. (2014). Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies. *Hum. Mol. Genet.* 23, R89–R98. doi:10.1093/hmg/ddu328

Delgado, J.-S., Vodonos, A., Delgado, B., Jotkowitz, A., Rosenthal, A., Fich, A., et al. (2012). Primary Biliary Cirrhosis in Southern Israel: A 20year Follow up Study. *Eur. J. Intern. Med.* 23, e193–e198. doi:10.1016/j.ejim.2012.09.004

Dong, S.-S., Zhang, K., Guo, Y., Ding, J.-M., Rong, Y., Feng, J.-C., et al. (2021). Phenome-Wide Investigation of the Causal Associations between Childhood BMI and Adult Trait Outcomes: a Two-Sample Mendelian Randomization Study. *Genome Med.* 13, 48. doi:10.1186/s13073-021-00865-3

dos Santos Silva, I., and Swerdlow, A. (1993). Thyroid Cancer Epidemiology in England and Wales: Time Trends and Geographical Distribution. *Br. J. Cancer* 67, 330–340. doi:10.1038/bjc.1993.61

Elbers, L. P. B., Kastelein, J. J. P., and Sjouke, B. (2016). Thyroid Hormone Mimetics: the Past, Current Status and Future Challenges. *Curr. Atheroscler. Rep.* 18, 14. doi:10.1007/s11883-016-0564-7

Ellervik, C., Roselli, C., Christophersen, I. E., Alonso, A., Pietzner, M., Sitlani, C. M., et al. (2019). Assessment of the Relationship between Genetic Determinants of Thyroid Function and Atrial Fibrillation. *JAMA Cardiol.* 4, 144–152. doi:10.1001/jamacardio.2018.4635

Elta, G. H., Sepersky, R. A., Goldberg, M. J., Connors, C. M., Miller, K. B., and Kaplan, M. M. (1983). Increased Incidence of Hypothyroidism in Primary Biliary Cirrhosis. *Dig. Dis Sci.* 28, 971–975. doi:10.1007/BF01311724

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian Randomization. *JAMA* 318, 1925–1926. doi:10.1001/jama.2017.17219

Evans, D. M., and Davey Smith, G. (2015). Mendelian Randomization: New Applications in the Coming Age of Hypothesis-free Causality. *Annu. Rev. Genom. Hum. Genet.* 16, 327–350. doi:10.1146/annurev-genom-090314-050016

Fenneman, A. C., Rampanelli, E., Yin, Y. S., Ames, J., Blaser, M. J., Fliers, E., et al. (2020). Gut Microbiota and Metabolites in the Pathogenesis of Endocrine Disease. *Biochem. Soc. Trans.* 48, 915–931. doi:10.1042/BST20190686

Ferrari, S. M., Fallahi, P., Elia, G., Ragusa, F., Ruffilli, I., Paparo, S. R., et al. (2020). Thyroid Autoimmune Disorders and Cancer. *Semin. Cancer Biol.* 64, 135–146. doi:10.1016/j.semcancer.2019.05.019

Floreani, A., Franceschet, I., Cazzagon, N., Spinazzè, A., Buja, A., Furlan, P., et al. (2015). Extrahepatic Autoimmune Conditions Associated with Primary Biliary Cirrhosis. *Clinic Rev. Allerg Immunol.* 48, 192–197. doi:10.1007/s12016-014-8427-x

Garber, J. R., Cobin, R. H., Gharib, H., Hennessey, J. V., Klein, I., Mechanick, J. I., et al. (2012). Clinical Practice Guidelines for Hypothyroidism in Adults: Cosponsored by the American Association of Clinical Endocrinologists and the American Thyroid Association. *Thyroid* 22, 1200–1235. doi:10.1089/thy.2012.0205

Gershwin, M. E., Selmi, C., Worman, H. J., Gold, E. B., Watnik, M., Utts, J., et al. (2005). Risk Factors and Comorbidities in Primary Biliary Cirrhosis: a Controlled Interview-Based Study of 1032 Patients. *Hepatology* 42, 1194–1202. doi:10.1002/hep.20907

Gilgenkrantz, H., and Collin de l'Hortet, A. (2018). Understanding Liver Regeneration: From Mechanisms to Regenerative Medicine. *Am. J. Pathol.* 188, 1316–1327. doi:10.1016/j.ajpath.2018.03.008

Glanville, K. P., Coleman, J. R. I., O'Reilly, P. F., Galloway, J., and Lewis, C. M. (2021). Investigating Pleiotropy between Depression and Autoimmune

Diseases Using the UK Biobank. *Biol. Psychiatry Glob. Open Sci.* 1, 48–58. doi:10.1016/j.bpsgos.2021.03.002

Hemani, G., Bowden, J., and Davey Smith, G. (2018). Evaluating the Potential Role of Pleiotropy in Mendelian Randomization Studies. *Hum. Mol. Genet.* 27, R195–R208. doi:10.1093/hmg/ddy163

Hernandez, B. Y., Rahman, M., Loo, L. W. M., Chan, O. T. M., Horio, D., Morita, S., et al. (2021). BRAFV600E, Hypothyroidism, and Human Relaxin in Thyroid Carcinogenesis. *J. Cancer Res. Clin. Oncol.* 147, 183–194. doi:10.1007/s00432-020-03401-9

Huang, M.-J., and Liaw, Y.-F. (1995). Clinical Associations between Thyroid and Liver Diseases. *J. Gastroenterol. Hepatol.* 10, 344–350. doi:10.1111/j.1440-1746.1995.tb01106.x

Lindor, K. D., Bowlus, C. L., Boyer, J., Levy, C., and Mayo, M. (2019). Primary Biliary Cholangitis: 2018 Practice Guidance from the American Association for the Study of Liver Diseases. *Hepatology* 69, 394–419. doi:10.1002/hep.30145

Lleo, A., Marzorati, S., Anaya, J.-M., and Gershwin, M. E. (2017). Primary Biliary Cholangitis: a Comprehensive Overview. *Hepatol. Int.* 11, 485–499. doi:10.1007/s12072-017-9830-1

Milette, S., Hashimoto, M., Perrino, S., Qi, S., Chen, M., Ham, B., et al. (2019). Sexual Dimorphism and the Role of Estrogen in the Immune Microenvironment of Liver Metastases. *Nat. Commun.* 10, 5745. doi:10.1038/s41467-019-13571-x

Noyce, A. J., Kia, D. A., Hemani, G., Nicolas, A., Price, T. R., De Pablo-Fernandez, E., et al. (2017). Estimating the Causal Influence of Body Mass index on Risk of Parkinson Disease: A Mendelian Randomisation Study. *Plos Med.* 14, e1002314. doi:10.1371/journal.pmed.1002314

Ong, J. S., and MacGregor, S. (2019). Implementing MR-PRESSO and GCTA-GSMR for Pleiotropy Assessment in Mendelian Randomization Studies from a Practitioner's Perspective. *Genet. Epidemiol.* 43, 609–616. doi:10.1002/gepi.22207

Pacini, F., Castagna, M. G., Brilli, L., and Pentheroudakis, G. (2012). Thyroid Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* 23 (Suppl. 7), vii110–vii119. doi:10.1093/annonc/mds230

Paknys, G., Kondrotas, A., and Kėvelaitis, E. (2009). Risk Factors and Pathogenesis of Hashimoto's Thyroiditis. *Medicina* 45, 574–583. doi:10.3390/medicina45070076

Patil, N., Rehman, A., and Jialal, I. (2021). *StatPearls: Hypothyroidism*. Treasure Island (FL): StatPearls Publishing.

Pratt, D. S. (2016). Primary Biliary Cholangitis--A New Name and a New Treatment. *N. Engl. J. Med.* 375, 685–687. doi:10.1056/NEJMe1607744

Sharma, A., Punekar, P., and Jain, A. (2018). A Study of Thyroid Dysfunction in Cirrhosis of Liver and Correlation with Severity of Liver Disease. *Indian J. Endocr. Metab.* 22, 645–650. doi:10.4103/ijem.IJEM_25_18

Qin, J., Li, L., Jin, Q., Guo, D., Liu, M., Fan, C., et al. (2018). Estrogen Receptor β Activation Stimulates the Development of Experimental Autoimmune Thyroiditis through Up-Regulation of Th17-type Responses. *Clin. Immunol.* 190, 41–52. doi:10.1016/j.clim.2018.02.006

Rashkin, S. R., Graff, R. E., Kachuri, L., Thai, K. K., Alexeeff, S. E., Blatchins, M. A., et al. (2020). Pan-cancer Study Detects Genetic Risk Variants and Shared Genetic Basis in Two Large Cohorts. *Nat. Commun.* 11, 4423. doi:10.1038/s41467-020-18246-6

Reyes, H., Levi, A. J., Gatmaitan, Z., and Arias, I. M. (1971). Studies of Y and Z, Two Hepatic Cytoplasmic Organic Anion-Binding Proteins: Effect of Drugs, Chemicals, Hormones, and Cholestasis. *J. Clin. Invest.* 50, 2242–2252. doi:10.1172/JCI106721

Reyna, B., and Pickler, R. (1999). Patterns of Genetic Inheritance. *Neonatal. Netw.* 18, 7–10. doi:10.1891/0730-0832.18.1.7

Rosa, R., Cristoferi, L., Tanaka, A., and Invernizzi, P. (2018). Geoepidemiology and (Epi-)genetics in Primary Biliary Cholangitis. *Best Pract. Res. Clin. Gastroenterol.* 34-35, 11–15. doi:10.1016/j.bpg.2018.05.011

Salata, R., Klein, I., and Levey, G. (1985). Thyroid Hormone Homeostasis and the Liver. *Semin. Liver Dis.* 5, 29–34. doi:10.1055/s-2008-1041755

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., et al. (2018). Genome-wide Association Meta-Analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence. *Nat. Genet.* 50, 912–919. doi:10.1038/s41588-018-0152-6

Schussler, G. C., Schaffner, F., and Korn, F. (1978). Increased Serum Thyroid Hormone Binding and Decreased Free Hormone in Chronic Active Liver Disease. *N. Engl. J. Med.* 299, 510–515. doi:10.1056/NEJM197809072991003

Silveira, M. G., Mendes, F. D., Diehl, N. N., Enders, F. T., and Lindor, K. D. (2009). Thyroid Dysfunction in Primary Biliary Cirrhosis, Primary Sclerosing Cholangitis and Non-alcoholic Fatty Liver Disease. *Liver Int.* 29, 1094–1100. doi:10.1111/j.1478-3231.2009.02003.x

Suzuki, Y., Ishida, K., Takahashi, H., Koeda, N., Kakisaka, K., Miyamoto, Y., et al. (2016). Primary Biliary Cirrhosis Associated with Graves' Disease in a Male Patient. *Clin. J. Gastroenterol.* 9, 99–103. doi:10.1007/s12328-016-0635-x

Taylor, P. N., Albrecht, D., Scholz, A., Gutierrez-Buey, G., Lazarus, J. H., Dayan, C. M., et al. (2018). Global Epidemiology of Hyperthyroidism and Hypothyroidism. *Nat. Rev. Endocrinol.* 14, 301–316. doi:10.1038/nrendo.2018.18

Teumer, A., Chaker, L., Groeneweg, S., Li, Y., Di Munno, C., Barbieri, C., et al. (2018). Genome-wide Analyses Identify a Role for SLC17A4 and AADAT in Thyroid Hormone Regulation. *Nat. Commun.* 9, 4455. doi:10.1038/s41467-018-06356-1

Vincken, S., Reynaert, H., Schiettecatte, J., Kaufman, L., and Velkeniers, B. (2017). Liver Cirrhosis and Thyroid Function: Friend or Foe? *Acta Clin. Belg.* 72, 85–90. doi:10.1080/17843286.2016.1215641

Wang, Y., Guo, P., Liu, L., Zhang, Y., Zeng, P., and Yuan, Z. (2021). Mendelian Randomization Highlights the Causal Role of Normal Thyroid Function on Blood Lipid Profiles. *Endocrinology* 162, bqab037. doi:10.1210/endocr/bqab037

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J. C., et al. (2019). A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits. *Nat. Genet.* 51, 1339–1348. doi:10.1038/s41588-019-0481-0

Younossi, Z. M., Bernstein, D., Shiffman, M. L., Kwo, P., Kim, W. R., Kowdley, K. V., et al. (2019). Diagnosis and Management of Primary Biliary Cholangitis. *Am. J. Gastroenterol.* 114, 48–63. doi:10.1038/s41395-018-0390-3

Yuan, S., Kar, S., Vithayathil, M., Carter, P., Mason, A. M., Burgess, S., et al. (2020). Causal Associations of Thyroid Function and Dysfunction with Overall, Breast and Thyroid Cancer: A Two-sample Mendelian Randomization Study. *Int. J. Cancer* 147, 1895–1903. doi:10.1002/ijc.32988

Zeng, P., and Zhou, X. (2019). Causal Effects of Blood Lipids on Amyotrophic Lateral Sclerosis: a Mendelian Randomization Study. *Hum. Mol. Genet.* 28, 688–697. doi:10.1093/hmg/ddy384

Zeng, P., Wang, T., Zheng, J., and Zhou, X. (2019). Causal Association of Type 2 Diabetes with Amyotrophic Lateral Sclerosis: New Evidence from Mendelian Randomization Using GWAS Summary Statistics. *BMC Med.* 17, 225. doi:10.1186/s12916-019-1448-9

Zeng, Q., Zhao, L., Wang, C., Gao, M., Han, X., Chen, C., et al. (2020). Relationship between Autoimmune Liver Disease and Autoimmune Thyroid Disease: a Cross-Sectional Study. *Scand. J. Gastroenterol.* 55, 216–221. doi:10.1080/00365521.2019.1710766

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 January 2022
doi: 10.3389/fgene.2021.779455

Identifying Potential miRNA Biomarkers for Gastric Cancer Diagnosis Using Machine Learning Variable Selection Approach

Neda Gilani[1]*, *Reza Arabi Belaghi*[2,3], *Younes Aftabi*[4], *Elnaz Faramarzi*[5], *Tuba Edgünlü*[6] and *Mohammad Hossein Somi*[5]

[1]Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran, [2]Department of Mathematics, Uppsala University, Uppsala, Sweden, [3]Department of Statistics, Faculty of Mathematical Science, University of Tabriz, Tabriz, Iran, [4]Tuberculosis and Lung Diseases Research Center, Tabriz University of Medical Sciences, Tabriz, Iran, [5]Liver and Gastrointestinal Diseases Research Center, Tabriz University of Medical Sciences, Tabriz, Iran, [6]Department of Medical Biology, Faculty of Medicine, Muğla Sıtkı Koçman University, Muğla, Turkey

Aim: This study aimed to accurately identification of potential miRNAs for gastric cancer (GC) diagnosis at the early stages of the disease.

Methods: We used GSE106817 data with 2,566 miRNAs to train the machine learning models. We used the Boruta machine learning variable selection approach to identify the strong miRNAs associated with GC in the training sample. We then validated the prediction models in the independent sample GSE113486 data. Finally, an ontological analysis was done on identified miRNAs to eliciting the relevant relationships.

Results: Of those 2,874 patients in the training the model, there were 115 (4%) patients with GC. Boruta identified 30 miRNAs as potential biomarkers for GC diagnosis and hsa-miR-1343-3p was at the highest ranking. All of the machine learning algorithms showed that using hsa-miR-1343-3p as a biomarker, GC can be predicted with very high precision (AUC; 100%, sensitivity; 100%, specificity; 100% ROC; 100%, Kappa; 100) using with the cut-off point of 8.2 for hsa-miR-1343-3p. Also, ontological analysis of 30 identified miRNAs approved their strong relationship with cancer associated genes and molecular events.

Conclusion: The hsa-miR-1343-3p could be introduced as a valuable target for studies on the GC diagnosis using reliable biomarkers.

Keywords: miRNA, machine learning, boruta algorithm, gastric cancer, hsa-miR-1343-3p, AUC, GSE106817, GSE113486

OPEN ACCESS

Edited by:
Tao Wang,
Northwestern Polytechnical University, China

Reviewed by:
Xiawei Li,
Zhejiang University, China
Yanshuo Chu,
University of Texas MD Anderson Cancer Center, United States

***Correspondence:**
Neda Gilani
gilanin@tbzmed.ac.ir

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 September 2021*
Accepted: *22 November 2021*
Published: *10 January 2022*

Citation:
Gilani N, Arabi Belaghi R, Aftabi Y, Faramarzi E, Edgünlü T and Somi MH (2022) Identifying Potential miRNA Biomarkers for Gastric Cancer Diagnosis Using Machine Learning Variable Selection Approach. Front. Genet. 12:779455. doi: 10.3389/fgene.2021.779455

INTRODUCTION

Gastric cancer (GC) is a significant global health issue due to being the fifth leading cancer worldwide as well as the third cancer-related death leading cause, which leads to nearly 8,00,000 deaths annually (Bray, 2018). Morbidity and mortality due to GC have reduced in recent years, though the rate of 5-year survival is still fairly low (Howlader, 2014). A significant prognostic factor is the stage of cancer at the diagnosis time. The 5-year survival of GC patients is below 30% if the disease is diagnosed at the advanced stages (Hundahl et al., 2000), while the 5-year survival of patients ranges between 70 and 90% if diagnosed at the early stages (Choi, 2015). Thus, GC will remain among the toughest

challenges for physicians and researchers for so long since GC is not symptomatic until the advanced stages; this is why effective screening approaches for the early detection of GC are mandatory to overcome GC mortalities (Penon et al., 2014). Presently, gastroscopy is yet the standard test to diagnose GC (Veitch et al., 2015). Nonetheless, this screening approach is invasive and costly. Furthermore, minimally invasive or non-invasive markers, including carcinoembryonic antigen (CEA) and carbohydrate antigen 19-9 (CA19-9) have been commonly used clinically, though these markers are neither specific nor sensitive enough for GC early diagnosis (Carpelan-Holmström et al., 2002). Due to non-specific symptoms and the absence of an early diagnosis, a great number of patients with GC are diagnosed at the advanced stages (Hundahl et al., 2000; Hartgrink et al., 2009). Thus, cost-effective and non-invasive biomarkers are immediately required for the early diagnosis of GC.

Recent genome analysis revealed several biomarkers which are related to RNA, DNA, exosome, et cetera. A class of endogenous non-coding RNAs is MicroRNAs (miRNAs) (nearly 22 nt) which module the expression of the gene after transcription through degradation or translation blockage of target mRNAs (Bartel, 2004; Caldas and Brenton, 2005). It is well-known that cancer cells may release miRNAs via exosomes to enhance proliferation and migration (Li, 2018; Yoshimura, 2018; Zeng, 2018). The exosomal miRNAs released into biofluids, including serum, plasma, tear, urine, and gastric juice, may escape being degraded by RNases (Gilad, 2008). Moreover, miRNAs have been suggested as potential biomarkers which may be used to diagnose several types of cancers, including testicular germ cell tumors (using miRNA-371a-3p: specificity 94.0% and sensitivity 90.1%) (Dieckmann, 2019), bladder cancer (using 7-miRNA panel: specificity 87% and sensitivity 95%) (Usuba, 2019a), and hepatocellular carcinoma (using miR-424: specificity 87.13% and sensitivity 95.12%) (Lin, 2015), and lung cancer (Aftabi, 2021). Moreover, several studies reported that numerous miRNAs might be potentially used as biomarkers for GC diagnosis (Zhou, 2010; Cui, 2013; Su, 2014). Nonetheless, most of the miRNA biomarkers are not developed using comprehensive data mining according to miRNA profiling, and even they lack proper external efficacy validation (Link and Kupcinskas, 2018; Wei, 2019). Instead, recently, Artificial intelligence Technology (AT) usage in the field of microarray Data has attracted more attention. The disadvantage of the conventional statistical models, including logistic regression, was that they excluded the possible interaction terms and highly correlated variables; thus, they might lose a part of useful information, which might decrease their accuracy, specifically in the case of high dimensional miRNA data analysis (Alpaydin, 2020). Furthermore, the traditional models are not able to capture variables' non-linear associations (James et al., 2013; Gilani et al., 2017; Gilani et al., 2019). Instead, Machine Learning (ML) is able to deal with non-linear structures as well as detecting all the possible interactions which may exist between predictors (Gilani et al., 2018; Wiemken and Kelley, 2020).

Machine learning has several algorithms of which the decision trees (DT), random forests (RF), extreme gradient boosted trees (XGBT), and artificial neural networks (ANN) that have been frequently applied in medicine (Cleophas and Zwinderman, 2015; Deo, 2015), particularly in prediction of cancer (DeGregory, 2018; Fakhari et al., 2019). Random forest is a tree-based classification algorithm, and as the name indicates, the algorithm creates a forest with a huge number of trees. It is an ensemble algorithm that combines multiple algorithms. The random forest creates a set of decision trees from a random sample of the training set. It repeats the process with multiple random samples and makes a final decision based on majority voting (Zhou, 2012). Briefly, gradient boosted trees combine multiple classification trees into an additively weighted classifier. Boosting refers to the method where sequentially ascertained trees were trained, meaning each observation was weighted by its error obtained by minimizing the appropriate loss of function in the previous iteration. In this way, boosting is a gradient descent algorithm (Christensen and Bastien, 2016) and forces the classifier to focus on aspects of the data that are difficult to learn (Hastie et al., 2009).

Artificial neural networks have been broadly used in medical studies (Darsey et al., 2015; DeGregory, 2018; Shahid et al., 2019). Such models perform satisfactorily, especially for classification problems with complex and non-linear associations between variables (Hastie et al., 2009). Briefly, artificial neural networks are based on a collection of artificial neurons, which receive and process inputs (predictors), transmit them to other artificial neurons, and produce an output (Zhou, 2012).

Considering the important role of GC early diagnosis in patient's survival rate and the lack of published article on identifying potential miRNAs for GC prediction at an early stage by AT, the present study aims to identify the potential miRNA for predicting GC by AT in the datasets of Gene Expression Omnibus (GEO) specifically with the stat of the art machine learning models. Traditional statistical models such as linear models previously has been used in looking for GC biomarkers and identified miRNAs with the potential prediction power (Yao, 2020), however, they have not implemented advanced methods such as machine learning and new variable selection approaches such as Synthetic Minority *Oversampling* Technique (SMOTE). In the present study, for the first time, we aimed to use those new techniques for identification of GC related miRNAs with a reliable cut-of and highest possible accurecy in the external validation.

METHODS

The Applied Datasets

For training sample, we used GSE106817 dataset that is available at https://www.ncbi.nlm.nih.gov/geo/. The dataset consist of the data of 2,566 miRNAs obtained from 2,759 non-cancer controls, and 115 GC cases (4%). In the original study the serum samples of cancer cases and non-cancer controls have been analyzed by microarray for miRNA expression profiles (Yokoi, 2018). For test sample we used GSE113486 dataset, which includes data of miRNA expression profiles from the serum samples of 40 GC cases (28.6%) and 100 normal controls (71.4%) (Usuba, 2019b). All the datasets were serum miRNA profiles based on the same microarray

TABLE 1 | Selected important miRNAs by Boruta Algorithm Using XGboost Algorithm.

No	miRNA	Importance	Se (%)	Sp (%)	PPV (%)	NPV (%)	AUC (%)	Accuracy (%)	Kappa (%)
1	hsa-miR-1343-3p	100.00	100.00	100.00	100.00	100.00	100.00	100.00	1.00
2	hsa-miR-1290	80.39	92.50	98.00	94.87	97.03	99.05	96.43	0.96
3	hsa-miR-5100	80.11	100.00	99.00	97.56	100.00	99.23	99.29	0.99
4	hsa-miR-6746-5p	64.57	100.00	93.00	85.11	100.00	97.23	95.00	0.95
5	hsa-miR-4532	64.85	67.50	100.00	100.00	88.50	95.11	90.71	0.91
6	hsa-miR-8073	61.79	97.50	100.00	100.00	99.01	100.00	99.29	0.99
7	hsa-miR-1228-5p	56.24	97.50	100.00	100.00	99.01	100.00	99.29	0.99
8	hsa-miR-1199-5p	54.12	62.50	97.00	89.29	86.61	92.56	87.14	0.87
9	hsa-miR-3622a-5p	54.49	80.00	99.00	96.97	92.52	97.26	93.57	0.94
10	hsa-miR-8060	53.75	85.00	98.00	94.44	94.23	98.79	94.29	0.94
11	hsa-miR-1246	50.42	92.50	100.00	100.00	97.09	99.90	97.86	0.98
12	hsa-miR-4787-3p	50.32	90.00	100.00	100.00	96.15	98.75	97.14	0.97
13	hsa-miR-6087	49.68	22.50	88.00	42.86	73.95	62.70	69.29	0.69
14	hsa-miR-4259	47.55	90.00	98.00	94.74	96.08	99.04	95.71	0.96
15	hsa-miR-6877-5p	46.90	92.50	94.00	86.05	96.91	97.73	93.57	0.94
16	hsa-miR-124-3p	45.42	92.50	94.00	86.05	96.91	96.81	93.57	0.94
17	hsa-miR-6787-5p	45.14	87.50	99.00	97.22	95.19	99.70	95.71	0.96
18	hsa-miR-4454	45.05	95.00	98.00	95.00	98.00	98.10	97.14	0.97
19	hsa-miR-6760-5p	45.42	90.00	94.00	85.71	95.92	98.58	92.86	0.93
20	hsa-miR-668-5p	45.24	77.50	98.00	93.94	91.59	96.44	92.14	0.92
21	hsa-miR-6762-5p	42.09	45.00	92.00	69.23	80.70	88.94	78.57	0.79
22	hsa-miR-3191-3p	40.43	75.00	94.00	83.33	90.38	93.48	88.57	0.89
23	hsa-miR-1268b	39.32	70.00	94.00	82.35	88.68	93.91	87.14	0.87
24	hsa-miR-1185-2-3p	39.13	30.00	87.00	48.00	75.65	53.88	70.71	0.71
25	hsa-miR-6131	38.30	87.50	98.00	94.59	95.15	99.21	95.00	0.95
26	hsa-miR-920	38.39	87.50	96.00	89.74	95.05	98.26	93.57	0.94
27	hsa-miR-4635	38.02	77.50	98.00	93.94	91.59	95.38	92.14	0.92
28	hsa-miR-6724-5p	37.28	45.00	81.00	48.65	78.64	74.35	70.71	0.71
29	hsa-miR-1185-1-3p	37.19	20.00	85.00	34.78	72.65	54.70	66.43	0.66
30	hsa-miR-422a	38.02	55.00	87.00	62.86	82.86	72.94	77.86	0.78

platform, 3D-Gene Huma miRNA V21_1.0.0 (39). The study was approved by the NCCH Institutional Review Board (2015-376, 2016-29) and the Research Ethics Committee of Medical Corporation Shintokai Yokohama Minoru Clinic (6019-18-3772). Written informed consent was obtained from each participant (42). This study was approved by the Ethics Committee of Tabriz University of Medical Sciences (No: IR. TBZMED.REC.1400.006).

Boruta Machine Learning Algorithm

We used the Boruta machine learning algorithm to select the most critical miRNAs related to GC in the training sample that produce the highest prediction accuracy. In short, Boruta selects the variables that have a high impact on the prediction accuracy by providing the "variable importance" (Kursa and Rudnicki, 2018). We used SMOTE random oversampling to balance the outcome in the GSE106817 data. We then used five-fold cross-validation to find the optimal hyper parameters on DT, RF, LR, XGBT, and ANN to choose the best approaches in the balanced sample using the most important variables selected by Boruta. Once the prediction models were developed, we applied them on the test sample GSE113486 to verify the accuracy of developed prediction approach. We looked for an algorithm that may generate a higher predictive power among the 5 ML algorithms in terms of the yielded areas under the ROC curves (AUCs). Sensitivity, specificity, positive predictive value, negative predictive value, misclassification rate, and Kappa were assessed. The guidelines of developing transparent multivariable prediction models was followed for these analysis (Moons, 2015).

GeneCodis Ontological Analysis

GeneCodis is a web-based tool for the ontological analysis of lists of genes, proteins, and regulatory elements like miRNAs, transcription factors, and CpGs. It can be used to determine biological annotations or combinations of annotations that are significantly associated to a list of genes under study with respect to a reference list. As well as single annotations, this tool allows users to simultaneously evaluate annotations from different sources, for example GO Biological Process and KEGG. To this end, and before computing *p*-values, it uses the apriori algorithm to extract sets of annotations that frequently co-occur in the analyzed list of genes (Garcia-Moreno, 2021). We used GeneCodis 4 (https://genecodis.genyo.es/) for ontological analysis of the identified miRNAs list.

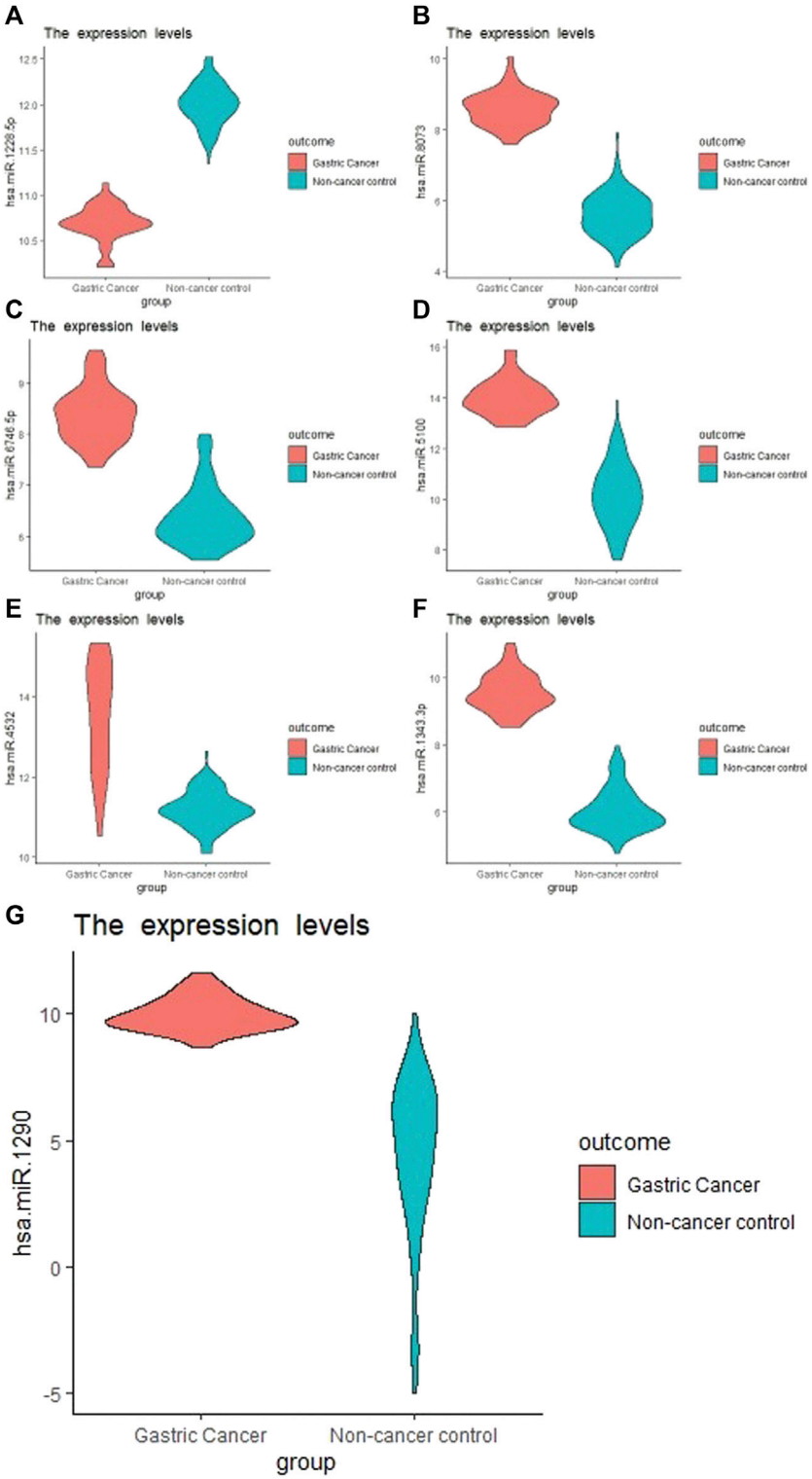

FIGURE 1 | Boxplot of the selected miRNA from Boruta Algorithm. **(A)**, hsa-miR-1228-5p; **(B)**, hsa-miR-8073; **(C)**, hsa-miR-6746-5p; **(D)**, hsa-miR-5100; **(E)**, hsa-miR-4532; **(F)**: hsa-miR-1343-3p; **(G)**, hsa-miR-1290.

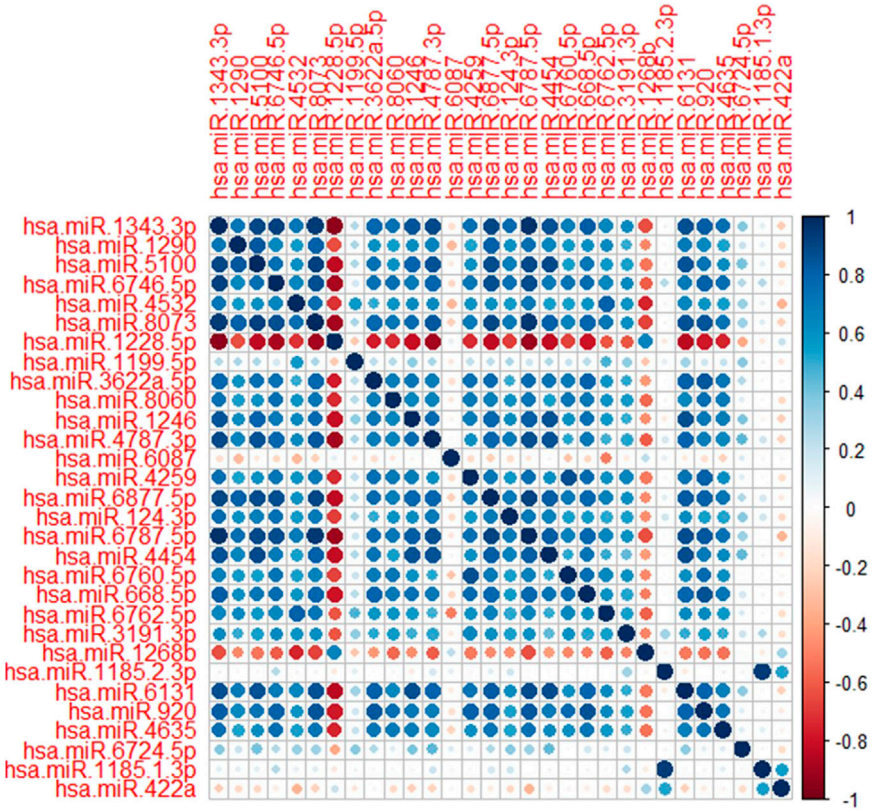

FIGURE 2 | Correlation plot of the selected miRNAs. Dark blue and dark red shows the strength of the correlations between miRNAs.

RESULTS

Of those 2,874 patients included in this study, there were 115 (4%) patients with gastric cancer. This analysis consists of 2,566 miRNAs.

Selected miRNAs as Potential GC Biomarkers

Of those 2,566 miRNA in GSE106817 data, the Boruta algorithm initially selected 108 miRNA using Gini Index measurement (results are not shown here). The processing time was 17.24 minuts There were 77 tentative variables at the first stage. After fixing the tentative features, Boruta identified 156 miRNA for the analysis (results are not shown here). The process took 99 iterations convergence. It was observed that hsa-miR-1343-3p had the highest importance for prediction accurecy (minimum importance; 6.47, median importance; 11.44, mean importance; 10.81; maximum importance; 13.63) among all identified miRNAs. The hsa-miR-1290 and hsa-miR-5100 had the second and third highest importance, with mean importance of 8.69 and 8.66, respectively (**Table 1**).

The balanced training data using SMOTE random oversampling technique had 1,376 cancer cases and 1,498 non-cancer controls. We trained DT, RF, LR, XGBT, and ANN perdition models with the selected miRNAs in the balanced training data.

Prediction Models and Accuracy in the Validation Data

The external validation data GSE113486 had 40 (28.6%) gastric cancer and 100 (71.4%) non-cancer (controls). hsa-miR-1343-3 produced the highest prediction accuracy for GC prediction (**Table 1**). For the hsa-miR-1343-3, all of the accuracy measures including AUC, sensitivity and specificity, positive predictive value, negative predictive value, Kappa were 100%. According to the decision trees, the cut-off point for this miRNA was 8.2 (**Figure 1**). Further, hsa-miR-8073 and hsa-miR-1228-5p produced 100% AUC but other accuracy measures were not 100%. On the other hsa-miR-1185-1-3p had the lowest AUC which has the least contribution to the prediction of GC.

Among several models discussed in the study, the XGBT algorithm had better prediction accuracy overall (Table S1-S4). However, for hsa-miR-1343-3 all models had consistently 100% accuracy which indicates that this miRNA may strongly predict GC. For some miRNA such as hsa-miR-422a XGBT algorithm could predict GC with higher accuracy than the logistic regression

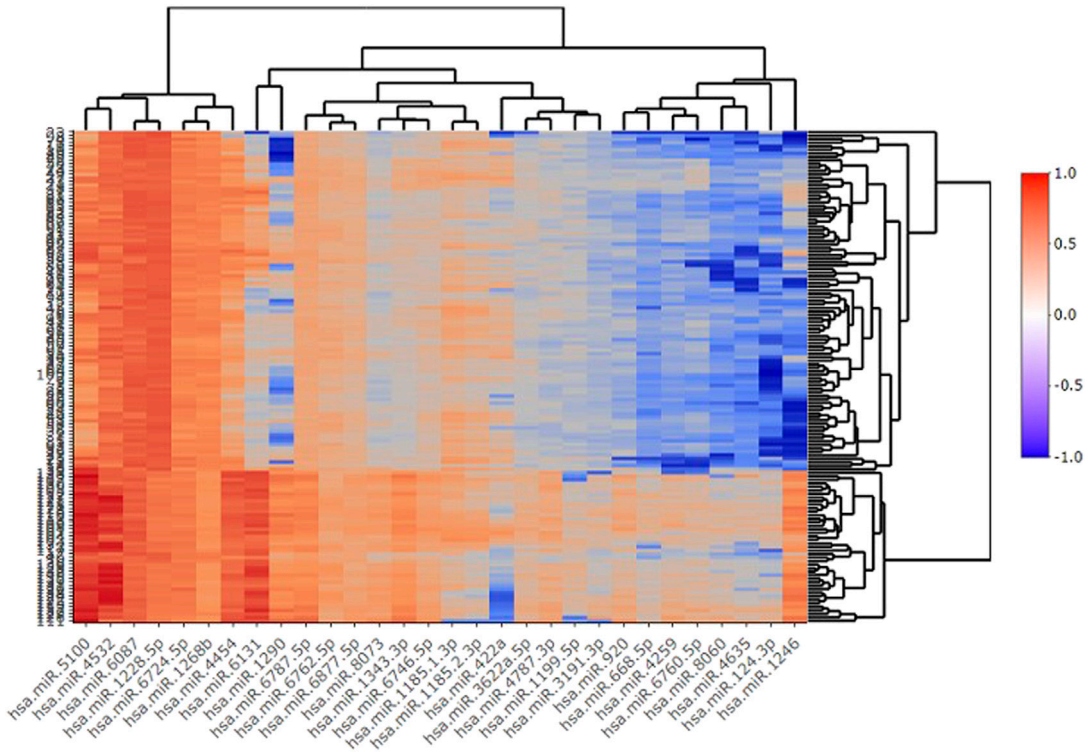

FIGURE 3 | Heathmap plot of clustering of 30 selected miRNAs.

and decision trees. **Figure 2** shows the correlation of the important miRNAs. It can be observed that most of the identified miRNAs except hsa-miR-422a, hsa-miR-1185-1-3p, hsa-miR-1185-2-3p, hsa-miR-6087, and hsa-miR-1199-5p are highly correlated. Consequently, clustering of correlated those miRNAs is helpful for the identification of cancerous and non-cancerous patients. Finally, Heatmap plot indicates the result of the hierarchical clustering analysis of the 30 selected miRNAs, which represents that identified miRNAs can easily distinguish GC cases and controls in test sample obtained from GSE113486 dataset (**Figure 3**).

Ontological Analysis

Regulatory, functional, and perturbation analysis by GeneCodis 4 showed that 30 identified miRNAs (**Table 1**) are related strongly to the cancer-associated genes and molecular events (**Figure 4**). Visualizations generated for 10 top terms of associations with Transcription Factors (**Figure 4A**), co-annotation of HMDD v3 (the Human microRNA disease database), MNDR (Mammalian ncRNA-Disease Repository), and TAM2 (The tool for annotations of human miRNAs) databases (**Figure 4B**), GO (Gene Ontology and GO Annotations) Biological Process (**Figure 4C**), GO Molecular Function (**Figure 4D**), co-annotation of KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathways, Panther (Protein ANalysis THrough Evolutionary Relationships) Pathways, and WikiPathways (**Figure 4E**), and co-annotation of HPO (The Human Phenotype Ontology) and OMIM (Online Mendelian Inheritance in Man) databases (**Figure 4F**).

DISCUSSION

Using artificial intelligence technology, we identified hsa-miR-1343-3 as a very strong nominate for biomarker analysis of GC diagnosis. The value of hsa-miR-1343-3 higher than 8.2 indicates that it could be a strong predictor for GC (100% of AUC, 100% of Sensitivity and Sepecificity). We also found three other miRNAs (hsa-miR-8073 and hsa-miR-1228-5p) with a great contribution to the GC prediction. A medical expert can use these findings for the early detection of GC instead of using costly and time-consuming tools such as colonoscopy Yao et *al.* (Yao, 2020).

This study had several strengths compared to the previous studies. Compared to Shi et al. that identified the miR-1246 as the potential biomarker of GC that generated the AUC of 83%, our study identified the hsa-miR-1343-3p using the Boruta algorithm that led to a significant increase in the AUC (Shi and Zhang, 2019). The study of Yao et al., selected three miRNAs that produced similar precision to our study that using only single miRNA that may have economical merits. Further, their study used a limited sample size (70 gastric cancer patients and 374 non-cancer controls) in the training set that may lead to an inferior model. The current study used very advanced variable selection methods and the state of the art machine learning approaches that produced consistent results. Another merit of

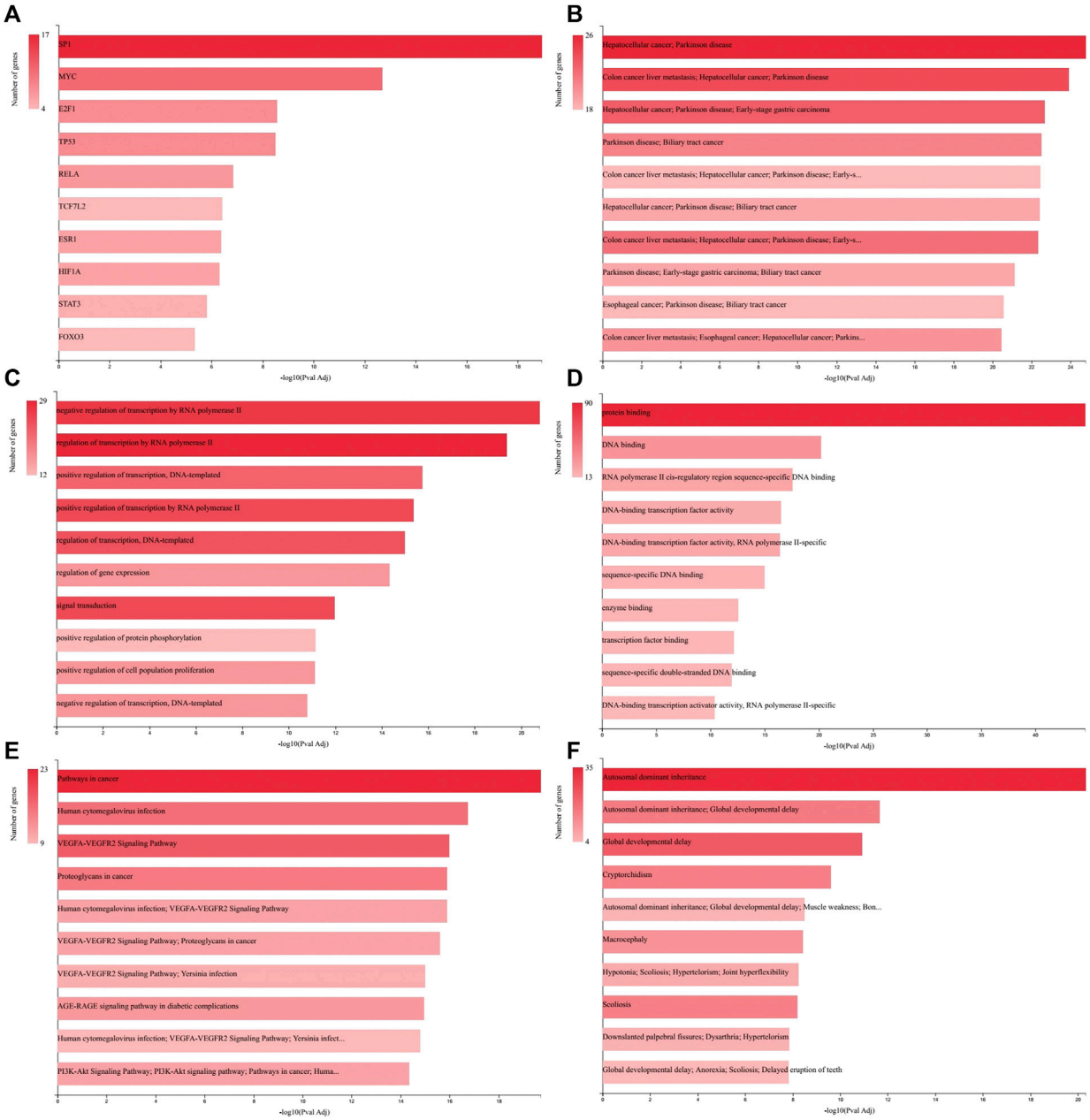

FIGURE 4 | GeneCodis Ontological analysis. Visualizations generated for 10 top terms of related categories with our identified miRNAs list are presented here for Transcription Factors **(A)**, Co-annotation of miRNAs-based analysis using HMDD v3, MNDR, and TAM2 **(B)**, GO Biological Process **(C)**, GO Molecular Function **(D)**, Co-annotation of KEGG Pathways, Panther Pathways, and WikiPathways databases **(E)**, and Co-annotation of HPO and OMIM databases **(F)**.

the study is introducing a simple cut-off point of 8.2 using decision trees that may has very practical value in GC classification.

Figure 4A depicted that among the transcription factors related to the genes associated with the identified miRNAs list (**Table 1**) the SP1, MYC, and E2F1 have higher priorities. SP1 protein expression is up regulated in GC tissues compared with normal tissues and is positively associated with depth of invasion and TNM stage of GC (Shi and Zhang, 2019). MYC is an oncogene responsible for excessive cell growth in cancer, enabling transcriptional activation of genes involved in cell cycle regulation, metabolism, and apoptosis, and is usually overexpressed in GC (Maués, 2018). E2F1 is a member of the E2F family that functions in cell cycle progression and apoptosis induction in response to DNA damage. Deregulated E2F1 acts as a driving force in GC progression and promotes tumor invasion and metastasis independently from its other cellular activities (Yan, 2014).

As depicted in **Figure 4B** gastrointestinal cancers including hepatocellular cancer, colon cancer, biliary tract cancer, and especially early-stage GC are among the most related diseases to the analyzed miRNAs list. From biological process and molecular function perspectives as showed in **Figures 4C,D**, the regulation of transcription and gene expression, and protein and DNA binding are the most targeted aspects, which are the general aspects of molecular biology of GC (Cervantes et al., 2007; Vauhkonen et al., 2006; Tan and Yeoh, 2015). Co-annotation of three pathway databases (**Figure 4E**) has shown that the miRNAs list is general in relation with pathways in cancer, VEGFA-VEGFR2 signaling pathway and PI3K-Akt signaling pathway. The increased expression of VEGFA in the tubular glands and VEGFR2 in the endothelium of GC samples mainly in the T2, T3 and T4 stages of tumor progression has been reported previously (Tamma, 2018). Also, it is showed that the PI3K/AKT/mTOR pathway is activated in GC with overexpression in tumor tissue, which is correlated with the depth of tumor infiltration and the presence of lymph node metastases (Tapia, 2014). Surprisingly, relation with Human cytomegalovirus infection, which was identified in our pathway analysis, has been reported to be associated with the development of GC (Jin, 2014) and GC lymphatic metastasis (Zhang, 2017).

The analysis of human phenotype and Mendelian inheritance ontologies identified Autosomal dominant inheritance and Global developmental delay among the most related phenomena with our miRNAs list. It is reported that gastric adenocarcinoma and proximal polyposis of the stomach is an autosomal dominant syndrome (Worthley, 2012). Also, some common variants have been described for GC and developmental delay (Hansford, 2015; Zhang, 2020).

In our study, we have shown theoreticaly that ther is a strong relationship between hsa-miR-1343-3p and GC. Hsa-miR-1343-3p has been indicated as a tumor suppressor for many types of cancer. It has been suggested that miR-1343-3p, which regulates the oncogenic effect of TEA domain transcription factors is associated with GC (Zhou, 2017). The correlation between hsa-miR-1343-3p and lung adenocarcinoma was evaluated and its expression was found to be low in patients with vascular invasion (Kim, 2017). Yuan et al. demonstrated that hsa-miR-1343-3p is consistently down-regulated in colon, prostate, and pancreatic cancers. Also, hsa-miR-1343-3p has been proposed as a biomarker to distinguish pancreato-biliary malignancy from non-malignant diseases. The major genes targeted by miR-1343-3p have been identified (Yuan, 2016). In this context, these target genes and their interaction with GC should also be investigated. The hsa-miR-1343-3p targets including SHISA7, TGFBR1, DLGAP3, SPRED1, ATXN7L3, and PLXDC2 genes are listed at MIRDB (http://mirdb.org/). Among them transforming growth factor beta-1 (TGFβ1) play an important role in carcinogenesis upon binding its receptor (TGFBR1). It acts as a tumor suppressor by inhibiting cellular proliferation or by promoting cellular differentiation and apoptosis. However, it turns to be a tumor promoter by stimulating angiogenesis and cell motility, suppressing the immune response, and increasing progressive invasion and metastasis (Yuan, 2016). Other reports have also revealed that hsa-miR-1343-3p reduces the expression of transforming growth factor-β (TGF-β) receptor-1, which induces angiogenesis through vascular endothelial growth factor (VEGF)-mediated apoptosis. Therefore, hsa-miR-1343-3p may also play an anti-angiogenic role (Ferrari et al., 2009; Stolzenburg et al., 2016; Kim, 2017). He et al. determined that TGFBR1 genes' two polymorphisms (rs334348, rs10512263) were associated with the risk of GC (He, 2018). In another study, Zhang et al. have shown that silencing of TGFBR1 inhibited cell proliferation, migration, invasion, and EMT in GC cells (Zhang, 2019).

Discs large associated proteins (DLGAPs) family has been implicated in psychological and neurological diseases. However, few studies have explored the association between the expression of DLGAPs and different types of cancer. Liu et al. has suggested that the significant overexpression of DLGAP4 in GC may be a promising potential prognostic marker for GC (Liu et al., 2018). Aslo, Liu et al. have determined decreased expression of SPRED1 in GC tissues (Liu et al., 2020).

However, there were certain limitations in our study. We had relatively small sample size in GC group. Other limitations were the pathological information such as the tumor stage, age or other factors which were not available in our datasets. Nonetheless, the prediction accuracy of our model has high enough (100% AUC) for clinical use. Further, we were unable to do the survival analysis to further validate the markers identified in this paper based on public available data (Howlader, 2014).

CONCLUSION

Using several state of the art machine learning methods and Boruta algorithm, we identified several miRNAs that can predict GC. Specifically, hsa-miR-1343-3p, which identified by cut-off point of 8.2 may be nominated as a highly reliable biomarker for, GC diagnosis after meticulous empirical tests.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

NG and RB conducted this study and performed statistical analysis and machine learning models. YA and TE provided ontological analysis. The paper was reviewed by EF and MS before final submission.

ACKNOWLEDGMENTS

We would like to thank the Associate Editor and three anonymous expert reviewers who significantly helped to improve the presentation of the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.779455/full#supplementary-material

REFERENCES

Aftabi, Y. (2021). Long Non-coding RNAs as Potential Biomarkers in the Prognosis and Diagnosis of Lung Cancer: A Review and Target Analysis. 73, 307–327. doi:10.1002/iub.2430

Alpaydin, E. (2020). *Introduction to Machine Learning*. MIT press.

Bartel, D. P. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *cell* 116, 281–297. doi:10.1016/s0092-8674(04)00045-5

Bray, F. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a Cancer J. clinicians* 68, 394–424. doi:10.3322/caac.21492

Caldas, C., and Brenton, J. D. (2005). Sizing up miRNAs as Cancer Genes. *Nat. Med.* 11, 712–714. doi:10.1038/nm0705-712

Carpelan-Holmström, M., Louhimo, J., Stenman, U.-H., Alfthan, H., and Haglund, C. C. E. A. (2002). CA 19-9 and CA 72-4 Improve the Diagnostic Accuracy in Gastrointestinal Cancers. *Anticancer Res.* 22, 2311–2316.

Cervantes, A., Braun, E. R., Fidalgo, A. P., González, I. C. J. C., and Oncology, T. (2007). *Mol. Biol. gastric Cancer* 9, 208–215. doi:10.1007/s12094-007-0041-4

Choi, I. J. (2015). Long-term Outcome Comparison of Endoscopic Resection and Surgery in Early Gastric Cancer Meeting the Absolute Indication for Endoscopic Resection. *Gastrointest. Endosc.* 81, 333–341. doi:10.1016/j.gie.2014.07.047

Christensen, J., and Bastien, C. (2016). Introduction to General Optimization Principles and Methods. *Nonlinear Optimization Vehicle Saf. Structures*, 107–168. doi:10.1016/b978-0-12-417297-5.00003-1

Cleophas, T. J., and Zwinderman, A. H. (2015). *Machine Learning in Medicine-A Complete Overview*. Springer.

Cui, L. (2013). Gastric Juice MicroRNAs as Potential Biomarkers for the Screening of Gastric Cancer. *Cancer* 119, 1618–1626. doi:10.1002/cncr.27903

Darsey, J. A., Griffin, W. O., Joginipelli, S., and Melapu, V. K. (2015). *Artificial Neural Networks 269-283*. Springer.

DeGregory, K. (2018). A Review of Machine Learning in Obesity. *Obes. Rev.* 19, 668–685. doi:10.1111/obr.12667

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation* 132, 1920–1930. doi:10.1161/circulationaha.115.001593

Dieckmann, K.-P. (2019). Serum Levels of microRNA-371a-3p (M371 Test) as a New Biomarker of Testicular Germ Cell Tumors: Results of a Prospective Multicentric Study. *J. Clin. Oncol.* 37, 1412. doi:10.1200/jco.18.01480

Fakhari, A., Gharepapagh, E., Dabiri, S., and Gilani, N. (2019). Correlation of Cancer Antigen 15-3 (CA15-3) Serum Level and Bony Metastases in Breast Cancer Patients. *Med. J. Islamic Republic Iran* 33, 142. doi:10.47176/mjiri.33.142

Ferrari, G., Cook, B. D., Terushkin, V., Pintucci, G., and Mignatti, P. (2009). Transforming Growth Factor-beta 1 (TGF-β1) Induces Angiogenesis through Vascular Endothelial Growth Factor (VEGF)-mediated Apoptosis. *J. Cell. Physiol.* 219, 449–458. doi:10.1002/jcp.21706

Garcia-Moreno, A. (2021). *GeneCodis 4: Expanding the Modular Enrichment Analysis to Regulatory Elements*.

Gilad, S. (2008). Serum microRNAs Are Promising Novel Biomarkers. *PloS one* 3, e3148. doi:10.1371/journal.pone.0003148

Gilani, N., Esmaeili, A., and Haghshenas, R. (2018). *The Effect of Eight Weeks Concurrent Training and Supplementation of L_Arginine on Plasma Level of 8-hydroxydeoxyguanosine (8-OHdG), Malondialdehyde and Total Antioxidant Capacity in Elderly Men (Multivariate Longitudinal Modeling)*.

Gilani, N., Haghshenas, R., and Esmaeili, M. (2019). Application of Multivariate Longitudinal Models in SIRT6, FBS, and BMI Analysis of the Elderly. *The Aging Male* 22, 260–265. doi:10.1080/13685538.2018.1477933

Gilani, N., Kazemnejad, A., Zayeri, F., Asghari, J. M., and Izadi, A. F. S. (2017). Predicting Outcomes in Traumatic Brain Injury Using the glasgow Coma Scale: A Joint Modeling of Longitudinal Measurements and Time to Event. *Iran Red Crescent Med J* 19 (2), e29663.

Hansford, S. (2015). *Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and beyond*. 1, 23–32.

Hartgrink, H. H., Jansen, E. P., van Grieken, N. C., and van de Velde, C. J. (2009). Gastric Cancer. *Lancet (London, England)* 374, 477–490. doi:10.1016/s0140-6736(09)60617-6

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Switzerland AG: Springer Science & Business Media.

He, B. (2018). Polymorphisms of TGFBR1, TLR4 Are Associated with Prognosis of Gastric Cancer in a Chinese Population. *Cancer Cel. Int.* 18, 1–10. doi:10.1186/s12935-018-0682-0

Howlader, N. (2014). *SEER Cancer Statistics Review, 1975–2011*, 19. Bethesda, MD: National Cancer Institute.

Hundahl, S. A., Phillips, J. L., and Menck, H. R. (2000). The National Cancer Data Base Report on Poor Survival of US Gastric Carcinoma Patients Treated with Gastrectomy: American Joint Committee on Cancer Staging, Proximal Disease, and the "Different Disease" Hypothesis. *Cancer* 88, 921–932. doi:10.1002/(sici)1097-0142(20000215)88:4<921:aid-cncr24>3.0.co;2-s

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Vol. 112. Springer.

Jin, J. (2014). *Latent Infection of Human Cytomegalovirus Is Associated with the Development of Gastric Cancer*, 8, 898–904. doi:10.3892/ol.2014.2148

Kim, H. (2017). MicroRNA Expression Profiles and Clinicopathological Implications in Lung Adenocarcinoma According to EGFR, KRAS, and ALK Status. *Oncotarget* 8, 8484. doi:10.18632/oncotarget.14298

Kursa, M. B., and Rudnicki, W. R. (2018). *Wrapper Algorithm for All Relevant Feature Selection*. Boruta: CRAN-Package.

Li, Z. (2018). Tumor-secreted Exosomal miR-222 Promotes Tumor Progression via Regulating P27 Expression and Re-localization in Pancreatic Cancer. *Cell Physiol. Biochem.* 51, 610–629. doi:10.1159/000495281

Lin, X.-J. (2015). A Serum microRNA Classifier for Early Detection of Hepatocellular Carcinoma: a Multicentre, Retrospective, Longitudinal Biomarker Identification Study with a Nested Case-Control Study. *Lancet Oncol.* 16, 804–815. doi:10.1016/s1470-2045(15)00048-0

Link, A., and Kupcinskas, J. (2018). MicroRNAs as Non-invasive Diagnostic Biomarkers for Gastric Cancer: Current Insights and Future Perspectives. *World J. Gastroenterol.* 24, 3313. doi:10.3748/wjg.v24.i30.3313

Liu, J., Liu, Z., Zhang, X., Gong, T., and Yao, D. (2018). Examination of the Expression and Prognostic Significance of DLGAPs in Gastric Cancer Using the TCGA Database and Bioinformatic Analysis. *Mol. Med. Rep.* 18, 5621–5629. doi:10.3892/mmr.2018.9574

Liu, W., Fang, S., and Zuo, G. (2020). *A Study on the Expression of SPRED1 and PBRM1 (Baf180) and Their Clinical Significances in Patients with Gastric Cancer*, 66. Clinical Laboratory. doi:10.7754/clin.lab.2020.200312

Maués, J. H. d. S. (2018). *Gastric Cancer Cell Lines Have Different MYC-Regulated Expression Patterns but Share a Common Core of Altered Genes*.

Moons, K. G. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann. Intern. Med.* 162, W1–W73. doi:10.7326/m14-0698

Penon, D., Cito, L., and Giordano, A. (2014). Novel Findings about Management of Gastric Cancer: a Summary from 10th IGCC. *World J. Gastroenterol. WJG* 20, 8986. doi:10.3748/wjg.v20.i27.8986

Shahid, N., Rappon, T., and Berta, W. (2019). Applications of Artificial Neural Networks in Health Care Organizational Decision-Making: A Scoping Review. *PloS one* 14, e0212356. doi:10.1371/journal.pone.0212356

Shi, S., and Zhang, Z. G. J. O. l. (2019). Role of Sp1 Expression in Gastric Cancer: A Meta-Analysis and Bioinformatics Analysis. 18, 4126–4135. doi:10.3892/ol.2019.10775

Stolzenburg, L. R., Wachtel, S., Dang, H., and Harris, A. (2016). miR-1343 Attenuates Pathways of Fibrosis by Targeting the TGF-β Receptors. *Biochem. J.* 473, 245–256. doi:10.1042/bj20150821

Su, Z.-X. (2014). Diagnostic and Prognostic Value of Circulating miR-18a in the Plasma of Patients with Gastric Cancer. *Tumor Biol.* 35, 12119–12125. doi:10.1007/s13277-014-2516-6

Tamma, R. (2018). *VEGFA and VEGFR2 RNAscope Determination in Gastric Cancer*, 49, 429–435. doi:10.1007/s10735-018-9777-0

Tan, P., and Yeoh, K.-G. J. G. (2015). *Genet. Mol. pathogenesis gastric adenocarcinoma* 149, 1153–1162. doi:10.1053/j.gastro.2015.05.059

Tapia, O. (2014). The PI3K/AKT/mTOR Pathway Is Activated in Gastric Cancer with Potential Prognostic and Predictive Significance. 465, 25–33. doi:10.1007/s00428-014-1588-4

Usuba, W. (2019). Circulating miRNA Panels for Specific and Early Detection in Bladder Cancer. *Cancer Sci.* 110, 408. doi:10.1111/cas.13856

Usuba, W. (2019). Circulating miRNA Panels for Specific and Early Detection in Bladder Cancer. 110, 408–419. doi:10.1111/cas.13856

Vauhkonen, M., Vauhkonen, H., Sipponen, P. J. B. P., and Gastroenterology, R. C. (2006). *Pathology and Molecular Biology of Gastric Cancer*, 20, 651–674. doi:10.1016/j.bpg.2006.03.016

Veitch, A. M., Uedo, N., Yao, K., and East, J. E. (2015). Optimizing Early Upper Gastrointestinal Cancer Detection at Endoscopy. *Nat. Rev. Gastroenterol. Hepatol.* 12, 660. doi:10.1038/nrgastro.2015.128

Wei, H. (2019). The Diagnostic Value of Circulating microRNAs as a Biomarker for Gastric Cancer: A Meta-Analysis. *Oncol. Rep.* 41, 87–102. doi:10.3892/or.2018.6782

Wiemken, T. L., and Kelley, R. R. (2020). Machine Learning in Epidemiology and Health Outcomes Research. *Annu. Rev. Public Health* 41, 21–36. doi:10.1146/annurev-publhealth-040119-094437

Worthley, D. (2012). Gastric Adenocarcinoma and Proximal Polyposis of the Stomach (GAPPS): a New Autosomal Dominant Syndrome. 61, 774–779. doi:10.1136/gutjnl-2011-300348

Yan, L.-H. (2014). Overexpression of E2F1 in Human Gastric Carcinoma Is Involved in Anti-cancer Drug Resistance. 14, 1–10. doi:10.1186/1471-2407-14-904

Yao, Y. (2020). Identification of Serum Circulating MicroRNAs as Novel Diagnostic Biomarkers of Gastric Cancer. *Front. Genet.* 11, 515. doi:10.3389/fgene.2020.591515

Yokoi, A. (2018). Integrated Extracellular microRNA Profiling for Ovarian Cancer Screening. *Nat. Commun.* 9, 1–10. doi:10.1038/s41467-018-06434-4

Yoshimura, A. (2018). Exosomal miR-99a-5p Is Elevated in Sera of Ovarian Cancer Patients and Promotes Cancer Cell Invasion by Increasing Fibronectin and Vitronectin Expression in Neighboring Peritoneal Mesothelial Cells. *BMC cancer* 18, 1–13. doi:10.1186/s12885-018-4974-5

Yuan, T. (2016). Plasma Extracellular RNA Profiles in Healthy and Cancer Patients. *Scientific Rep.* 6, 1–11. doi:10.1038/srep19413

Zeng, Z. (2018). Cancer-derived Exosomal miR-25-3p Promotes Pre-metastatic Niche Formation by Inducing Vascular Permeability and Angiogenesis. *Nat. Commun.* 9, 1–14. doi:10.1038/s41467-018-07810-w

Zhang, C. (2020). YARS as an Oncogenic Protein that Promotes Gastric Cancer Progression through Activating PI3K-Akt Signaling. 146, 329–342. doi:10.1007/s00432-019-03115-7

Zhang, L. (2019). Circular RNA CircCACTIN Promotes Gastric Cancer Progression by Sponging MiR-331-3p and Regulating TGFBR1 Expression. *Int. J. Biol. Sci.* 15, 1091. doi:10.7150/ijbs.31533

Zhang, L. (2017). Human Cytomegalovirus Detection in Gastric Cancer and its Possible Association with Lymphatic Metastasis. 88, 62–68. doi:10.1016/j.diagmicrobio.2017.02.001

Zhou, H. (2010). Detection of Circulating Tumor Cells in Peripheral Blood from Patients with Gastric Cancer Using microRNA as a Marker. *J. Mol. Med.* 88, 709–717. doi:10.1007/s00109-010-0617-2

Zhou, Y. (2017). TEAD1/4 Exerts Oncogenic Role and Is Negatively Regulated by miR-4269 in Gastric Tumorigenesis. *Oncogene* 36, 6518–6530. doi:10.1038/onc.2017.257

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 04 February 2022
doi: 10.3389/fgene.2022.828786

Integrative OMICS Data-Driven Procedure Using a Derivatized Meta-Analysis Approach

Karla Cervantes-Gracia[1], *Richard Chahwan*[1]* *and Holger Husi*[2,3]*

[1]*Institute of Experimental Immunology, University of Zurich, Zurich, Switzerland,* [2]*Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, United Kingdom,* [3]*Division of Biomedical Sciences, Centre for Health Science, University of the Highlands and Islands, Inverness, United Kingdom*

The wealth of high-throughput data has opened up new opportunities to analyze and describe biological processes at higher resolution, ultimately leading to a significant acceleration of scientific output using high-throughput data from the different omics layers and the generation of databases to store and report raw datasets. The great variability among the techniques and the heterogeneous methodologies used to produce this data have placed meta-analysis methods as one of the approaches of choice to correlate the resultant large-scale datasets from different research groups. Through multi-study meta-analyses, it is possible to generate results with greater statistical power compared to individual analyses. Gene signatures, biomarkers and pathways that provide new insights of a phenotype of interest have been identified by the analysis of large-scale datasets in several fields of science. However, despite all the efforts, a standardized regulation to report large-scale data and to identify the molecular targets and signaling networks is still lacking. Integrative analyses have also been introduced as complementation and augmentation for meta-analysis methodologies to generate novel hypotheses. Currently, there is no universal method established and the different methods available follow different purposes. Herein we describe a new unifying, scalable and straightforward methodology to meta-analyze different omics outputs, but also to integrate the significant outcomes into novel pathways describing biological processes of interest. The significance of using proper molecular identifiers is highlighted as well as the potential to further correlate molecules from different regulatory levels. To show the methodology's potential, a set of transcriptomic datasets are meta-analyzed as an example.

Keywords: meta-analysis, omics, bioinformatics, biomarker analysis, pathway analysis, data integration

OPEN ACCESS

Edited by:
Tao Wang,
Northwestern Polytechnical University, China

Reviewed by:
Yanshuo Chu,
University of Texas MD Anderson Cancer Center, United States
Kai Wang,
University of Iowa, United States

***Correspondence:**
Richard Chahwan
chahwan@immunology.uzh.ch
Holger Husi
Holger.Husi@uhi.ac.uk

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *03 December 2021*
Accepted: *12 January 2022*
Published: *04 February 2022*

Citation:
Cervantes-Gracia K, Chahwan R and Husi H (2022) Integrative OMICS Data-Driven Procedure Using a Derivatized Meta-Analysis Approach. Front. Genet. 13:828786. doi: 10.3389/fgene.2022.828786

1 INTRODUCTION

Traditional data analytical approaches focus on hypothesis-driven methods to understand specific and known molecular targets. Alternatively, data-driven approaches are based on high-throughput methodologies that provide un-biased genome-wide analysis of multiple omics variables which mirrors the different layers of biological regulation of a system. Undoubtedly, knowledge generated by traditional approaches through the years is essential to contextualize and properly analyze high-throughput data (McDermott et al., 2013; Guan et al., 2020). Ultimately, data-driven approaches aim to provide a number of potential hypotheses that feed into the traditional approach cycle in order to be validated or refuted (Fernandes and Husi 2019).

Nowadays, the surge of studies based on high-throughput data analysis has led to an expansion of public repositories (i.e., GEO, ArrayExpress) that store and provide access to these data for further analyses (Clough and Barrett 2016; Athar et al., 2019). As a consequence, big data production and availability have provided novel venues/opportunities for data interpretation, data integration, statistical analysis, and therefore new hypotheses that might reveal new inferences and provide a higher molecular resolution of a determined phenotype or disease. Nevertheless, the lack of a unified system to publish the different omics data generated and to report, curate and consolidate all the different identifiers available remains a challenge (Durinck et al., 2005; McGarvey et al., 2019). Unique outcomes generated from the different high-throughput technologies and the lack of standardized approaches to analyze, integrate and interpret these heterogeneous and often incompatible data have led to the emergence of different analytic methodologies that focus on varying ways of data-interpretation.

Meta-analytic methodologies have been commonly followed in data science to collate and identify commonalities across different studies, and to rule out inconsistencies commonly found in published literature (Waldron and Riester 2016; Vennou et al., 2020). The statistical basis of these methodologies provides valuable results and gives strength to the variables that reflect an association and consistency across studies. These methodologies are based on the fact that even amongst heterogeneous studies, associations can be made. Thus, meta-analysis can lead to the identification of robust and quantifiable variables shared across studies published by different groups–despite inherent differences in such cohorts–generated through different platforms and techniques that could have been otherwise overlooked (Care et al., 2015; Cho et al., 2016; Piras et al., 2019; Winter et al., 2019). In the big-data field, the exponential growth of high-throughput data availability has highlighted the advantage to follow meta-analysis methodologies in order to increase the statistical power of the outcomes and make sense out of the great amount of data shared within the scientific community (Xia et al., 2013; Kim et al., 2017; Forero 2019; Jaiswal et al., 2020; Vennou et al., 2020).

The generation and analysis of high-throughput data are commonly focused on a single biological parameter (e.g., transcripts, proteins, or metabolites) and represent only a snapshot of what is happening in a specific molecular process. Due to the high density of available studies, several meta-analytic approaches have been developed and standardized to integrate transcriptomic data. Effect size (t-statistic combination), rank-ratio (fold-change ratio combination), Fisher's (*p*-value combination), and vote-counting (VCS–number of reporting studies) are some of the common methods followed to perform a meta-analysis on these samples (Rikke et al., 2015; Goveia et al., 2016; Forero 2019; Shafi et al., 2019; Toro-Domínguez et al., 2020). Among the many promising applications of these approaches two stand out; namely, biomarker discovery and signaling pathway identification. The premise that biomarkers identified with computational approaches from a single high-throughput study exhibit little overlap with other studies indicates that these might represent false positives and cannot be fully trusted. Thus, meta-analyses have been long-performed with the goal to discover novel and robust biomarkers, distinguishable and consistent patterns of disease-associated deregulated genes. Statistically significant deregulated genes have been associated with several cancers and other diseases through the application of different meta-analytic approaches (Fishel et al., 2007; Xu et al., 2009; Huan et al., 2015; Cho et al., 2016; Bell et al., 2017; Piras et al., 2019; Su et al., 2019). Pathway analyses have also dominated the meta-analysis studies aiming to highlight the main deregulated processes to some extent (Kröger et al., 2016; Wang et al., 2017; Badr and Häcker 2019).

Nevertheless, due to the dynamicity of biological systems and the known crosstalk among the multiple layers of biological regulation, the orchestrated analysis of the different omics levels remains essential. Thus, the study of deregulated pathways and the implementation of integrative systems biology approaches seems logical and sought after techniques (Auffray et al., 2010; Norris et al., 2017; Parker et al., 2019; Shafi et al., 2019; Myall et al., 2021). These approaches have been highlighted by their potential to better understand the complex, albeit inevitable, interactions among different omics data. Ultimately, systems analysis aims to elucidate the regulation of pathways that might underpin cause and effect factors and improve the understanding of systems behavior by providing more accurate models of a determined condition of interest.

Although integrative systems biology approaches have been applied to individual studies by performing a variety of high-throughput omics approaches and analyzing multiple layers of gene regulation data (genetic variants, RNA transcripts, DNA methylation profiles, protein concentrations, chromatin marks) (Boeing et al., 2016; Saha et al., 2018; Xicota et al., 2019; Mair et al., 2020), the possibility to sum-up and analyze publicly available data generated by different scientific groups from individual omics approaches through multi-study meta-analyses may not only increase the statistical power of the outcomes but enhance and complement the biological knowledge through the re-analysis and integration of large-scale data; thereby highlighting significant but previously undetectable molecular links.

Various methodologies are available to pursue systems biology analyses, each of which follows different strategies, with associated limitations and outcomes (**Table 1**) (Xia et al., 2013; Rohart et al., 2017; Argelaguet et al., 2018; Forero 2019; Singh et al., 2019; Winter et al., 2019; Zhou et al., 2019; Pang et al., 2020; Toro-Domínguez et al., 2020; Yang 2020; Zhou et al., 2020). Here we aim to describe the Harmonized Holistic (HH) meta method, a simplistic, flexible, adjustable, and scalable methodology (limited only by the availability of omics data) that can go from single omics to multi-omics analyses (**Figure 1**). Our methodology is not based on a computational approach, it is a meta-analysis based on case and control comparisons of pre-processed data per study. The basis of the data to perform integrative approaches is of importance, and there is where this meta-analysis approach gears towards allowing heterogeneous omics data integration. It can integrate unmatched mRNA, miRNA, DNA methylation profiles, protein, metabolites

TABLE 1 | Comparison of available integrative systems biology methodologies.

Methodology	Strategy	Outcome	Limitations	References
HHmeta method	Meta-analysis of Differentially Expressed molecules from omics data. Data from different platforms (e.g. RNAseq, microarray) can be integrated. Biomarker list generation by ranking the frequency distribution and contextualization of molecules into pathways	- Integration of omics Biomarker lists and contextualization into pathway maps - Novel hypotheses from the Biomarker list - Novel hypotheses from the Main deregulated pathways - Better understanding of the disease/condition of interest	- Depends on availability of the data - Relies on previous knowledge - Gaps prevail across the pathway maps - Molecules without a defined function or interaction are not mapped	Cervantes-Gracia et al. (2021); current paper
Network meta-analysis	Meta-analysis of transcriptomics data by including Differentially Expressed comparison analysis per independent study	- Differentially Expressed Gene list based on meta-analysis of independent experimental studies - GSEA.	- Do not focuses on integrate different omics - Focus on obtaining signatures/ biomarkers	Winter et al. (2019)
MetaPCA	Meta-analysis of transcriptomic or epigenomic datasets through identification of a common eigen-space for dimension reduction	- Clusters and Patterns of gene expression profile - Robust to outliers	- Do not focuses on integrate different omics - Focus on obtaining signatures/ molecular patterns	Kim et al. (2018)
MINT	Independent omics studies integration based on similar biological questions Allows supervised and unsupervised frameworks. It is a PLS-based method to model multi-group (studies) data	- Identification of reproducible biomarker signatures	- It can only include studies with a sample size bigger than 3 - Focus on obtaining signatures/ biomarkers	Rohart et al. (2017)
NetworkAnalyst	Gene expression profiling, meta-analysis and systems-level interpretation	- Creates and visualizes biological networks - Web-based meta-analysis of gene expression data - Comparison of multi gene lists generated outside the tool - Identification of shared and unique genes and processes, through multi-list heatmaps and enrichment networks	- Format of gene expression profiles outside the application - Integration of transcriptomics studies	Zhou et al. (2019)
Mergeomics	Multi-omics association data, pathway analysis and functional genomics, analysis. It corrects for dependencies between omics markers. Based on pathway or network-level meta-analysis	- Identification of key drivers of a disease and causal subnetworks for specific conditions - Single dataset: causal network or key regulatory genes can be identified - Multiple dataset (same or different data type): meta-analysis, causal networks, key regulatory genes - Groups of disease associated genes: key regulators, condition sub-networks, gene sets association with other conditions or organisms	- Format of gene expression profiles outside the application - Based on comparison files: Cases vs controls	Arneson et al. (2016)
INMEX	Meta-analysis of multiple gene-expression datasets that allows integration of transcriptomics and metabolomics datasets	- Data preparation - Statistical analysis: multiple datasets combination based on *p*-values, effect sizes, rank orders and other features - Functional analysis and ID combination between genes and metabolites	- Limited to integration of transcriptomics and metabolomics	Xia et al. (2013)
DIABLO	Multi-omics integrative, holistic and data-driven method	- Identification of known and novel multi-omics biomarkers - Identify correlated variables within omics datasets from the same samples	- Batch effect analyses in each dataset are needed prior to integration - Integration of different omics dataset from the same biological samples - Focus on obtaining signatures/ biomarkers	Singh et al. (2019)
MOFA	Unsupervised identification of principal sources of variation among multi-omics datasets	- Identification of factors specific to data modalities and common within multiple molecular layers	- Analysis and integration of different omics datasets from the same biological samples. Similar to DIABLO, JIVE, PARADIGM or MCIA.	Argelaguet et al. (2018); (Subramanian et al., 2020)

(Continued on following page)

TABLE 1 | (*Continued*) Comparison of available integrative systems biology methodologies.

Methodology	Strategy	Outcome	Limitations	References
Ingenuity pathway analysis (IPA)	Multi-omics pathway analysis tool	- Building of networks to represent biological systems - Pathway analysis and association of processes activation or inhibition in a specific condition - Identification of novel targets - Comparison across multiple analyses. Similar to Pathway studio (Elsevier)	- Linear model, thus, non-linear associations might be missed - Commercial - Do not generate meta-analyses - Un-reproducible results - Based on computational approaches	Ingenuity Pathway Analysis tool (IPA; QIAGEN Inc., Germantown, MD, USA, https://www.qiagenbioinformatics.com/products

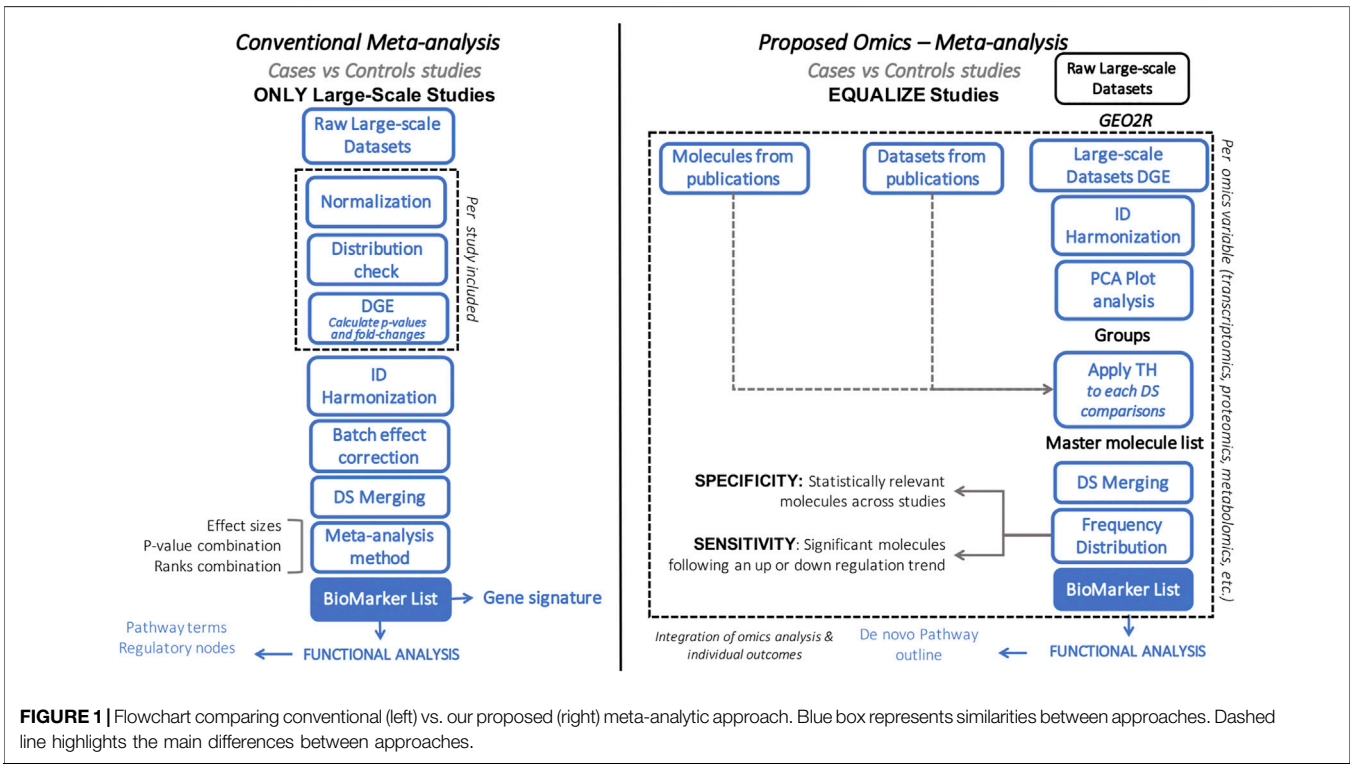

FIGURE 1 | Flowchart comparing conventional (left) vs. our proposed (right) meta-analytic approach. Blue box represents similarities between approaches. Dashed line highlights the main differences between approaches.

information from independent studies that meet the inclusion criteria of a specific research question. By considering the commonalities of the differentially expressed molecules across studies, the HHmeta method circumvents the variable depth of data produced by different measurement technologies (i.e., Microarray, RNAseq), as well as by high and low-throughput studies. This approach provides a ranking system that goes beyond the *p*-value and log2-Fold Change significance filtering, by defining the molecules with a significant and consistent trend in regulation among the different studies analyzed. Ultimately, this approach leads to the identification of pathways that can be fed and confirmed with the different omics data analyzed, which validates the outcomes and increase the significance of the identified targets (see Graphical Abstract in **Supplementary Presentation S1**).

We have previously explored our methodology in several iterations and proved its potential in a variety of disease settings (Fernandes and Husi 2016; Cervantes-Gracia and Husi 2018; Fernandes et al., 2018), however the ranking system was not properly established. Here, we 1) consolidate the final optimized pipeline and 2) apply this framework to 6 DLBCL (Diffuse Large B Cell Lymphoma) datasets (DS) from different studies and sources (tumoral tissue vs. b-cells). We aimed to identify the DS that indeed showed a potential correlation to further identify altered pathways coming only from B-cell deregulation (Cervantes-Gracia et al., 2021; Sheppard et al., 2018). The identified pathways represent the common deregulated processes found within the different DS included in the downstream analysis and delimit their significance by the outlined trend of the pathway identified. The outlined

pathways can be further fed with different omics data by complementing with the variables that correlate with the outcomes from the different omics levels. This analysis shows the potential of the presented methodology to not only identify potential biomarkers but also deregulated processes with a notable trend providing data-driven hypotheses that have either already been validated or that better yet have not been associated with the disease and need further validation.

2 MATERIALS AND METHODS

The cornerstone of this methodology is for each group of interest to be compared to their most appropriate controls. Hence, the integration of large-scale DS from a variety of publications relies on keeping the study groups per publication intact. Thus, the basis of the methodology is set on statistically significant molecules and their ratio-metric values (e.g., log-fold change) per comparison, identifiers curation through a unifier database, data-format and structure, outlier detection, as well as group re-stratification. This methodology has been previously performed manually (Cervantes-Gracia and Husi 2018). The aim of this work is to describe and summarize the whole procedure into a formula that statistically ranks and explains the significance of the molecules included in the biomarker list. Here, both the manual approach and the frequency score (FS) index are presented.

The particular molecules under study (e.g., miRs, mRNA, proteins) are individually processed and meta-analyzed before the multi-omics integration and analysis is performed. In this methodology, each molecular type has its own comparison matrix where the large-scale studies are merged. Thus, every individual DS comparison (case vs. control) has the main molecules included in the analysis facing each other. The output of the HHmeta method is divided on biomarker identification and functional analysis. The methodology outline is divided into four main sections: Data collection, Data correlation and structure, Grouping and Biomarker discovery and Data integration and Functional analysis.

2.1 Data Collection

High-throughput DS from publications can be collected from public repositories such as GEO NCBI, ExpressionAtlas and ArrayExpress from EMBL-EBI databases for raw and processed omics data and SRA for raw sequencing data. Specialized databases exist for the different omics data. PRIDE, Peptide Atlas, ProteomicsDB, GPMDB, JPOST repository, MassIVE, PAXDB for proteomics, MetaboLights, MetabolomeExpress, MetabolomicsWorkbench, GNPS for metabolomics and EGA, EVA for genomics are examples of available omics databases (Carroll et al., 2010; Deutsch 2010; Coutant et al., 2012; Wang et al., 2012; Vizcaíno et al., 2013; Fenyö and Beavis 2015; Lappalainen et al., 2015; Kale et al., 2016; Sud et al., 2016; Wang et al., 2016; Wang et al., 2018; Samaras et al., 2020; Watanabe et al., 2021). Platforms like Omics discovery index (OmicsDI) exist, where biological and technical metadata from public omics datasets are stored and standardized through an indexing system to enable access, discovery and broadcasting of omics datasets (Perez-Riverol et al., 2017). In terms of cancer databases TCGA, COSMIC, OCCPR and ICGC are distinguished high-throughput data repositories. Data can also be collected directly from the literature. The DS collected can be derived from entirely unmatched sources (e.g., DNA, RNA, protein), different platforms (e.g., Microarray, RNAseq), and samples (e.g., tissue, blood, urine).

Here, the example shows the analysis of DLBCL and the potential to correlate and find the common and significant molecules and deregulated mechanisms across expression profiles from tumoral vs healthy samples. The following DS (gene expression profiles) were retrieved from GEO (NCBI) database (Clough and Barrett 2016): GSE9327 (tumoral tissue vs healthy tissue; CNIO Human Oncochip), GSE32018 (tumoral tissue vs. healthy tissue; Agilent), GSE56315 (tumoral tissue vs. healthy B-cells; Affymetrix), GSE12195 (tumoral tissue vs. healthy B-cells; Affymetrix), GSE2350 (tumoral B-cells vs. healthy B-cells; Affymetrix), GSE12453 (tumoral B-cells vs. healthy B-cells; Affymetrix).

2.2 Data Correlation and Structure

This module comprises three steps: DS group comparison, Data ID harmonization, and Data Merging within and across DS comparisons.

2.2.1 Datasets Group Comparison

This step represents the first statistical evaluation embedded within this methodology. Here we rely on pre-processed and normalized available DS; raw data can also be considered. Raw samples need to be normalized individually to be further statistically assessed and generate ratio-metric values. Differential expression analysis of the GEO DS collected are performed through GEO2R, a web-based tool that includes GEO Query and Limma Bioconductor packages and performs multiple-testing correction through Benjamini–Hochberg false discovery rate method as a default (Benjamini and Hochberg 1995; Gentleman et al., 2004; Smyth 2004; Sean and Meltzer 2007). Data collected directly from the literature already provides ratio-metric values to integrate into the correlation matrix.

2.2.2 Data ID Harmonization

Given that different experimental platforms (e.g., different microarray technologies, RNAseq) and functional analysis tools usually produce and require unique identifiers, there is a need for standard names for each type of molecule (e.g., transcript, protein) under study. In order to be able to correlate the ratio-metric values of the molecules shared across every DS comparison included, and reduce data redundancy within studies, the DS needs to be mapped to a common identifier (e.g., Uniprot or PADB identifiers). Furthermore, a unifier that consolidates the different accession numbers and identifiers can be of great assistance. The PADB database was established by H. Husi (Husi 2004) as a unifier database for molecular data and it has been the reference database for several of our studies (available by request). PADB has been continuously curated and updated over the last 20 years. It contains old and recent

identifiers from multiple databases and platforms that have been assigned to the molecules through time. This database clusters identifiers from a variety of databases and platforms (Ensemble, Genenames, RefSeq, Uniprot, Swissprot, Agilent, Affymetrix, Illumina, and others), and provides a unique unifier ID that maps the molecules to all these reliable identifiers, allowing further data merging and analysis through a variety of tools. Uniprot database and BioMart also offer the option to retrieve alternative identifiers (e.g., Ensemble, Genenames) for molecules of interest or by downloading the complete file to index by any cross-indexing tool (Smedley et al., 2009; Consortium et al., 2021).

To cross-index and further merging of accession numbers, the in-house software AWASH was used. It is a text manipulation software that performs data cleaning and merging by using either a single file or multiple files, the latter based on a parent file as a reference for further indexing on dependent files. To perform the indexing, a Master file (containing the identifiers from the common database chosen and the accession number of the DS in question) and a Child file (different files for each DS comparison containing all accession numbers, statistics, and ratio-metric values) are needed. The input files should be in Tab- Separated Values (TSV) format. After indexing, each Child file accession number will be associated with a common unifier ID and alternative identifiers.

2.2.3 Data Merging Within and Across Datasets Comparisons

Often within large-scale DS, there is more than one probe-set and values for the same molecule. Thus, the significant (e.g., FDR/*p*-value) and ratio-metric (e.g., Fold change) values from molecules with the same unifier ID can be either merged or one can keep only the probe-sets that have the most significant values, as long as the same method is followed for each DS comparison merging. Every DS comparison should only contain one ID for each of the molecules within it. AWASH software can be used for this purpose based on the common unifier IDs and a Masterfile containing each of the DS comparisons.

Once all cleaned, all the DS comparisons from the same molecular type are merged into a matrix based on a list of unifiers reported among all the DS comparisons. The identifiers from all DS comparisons would follow the same order. Thus, the same molecule would be facing each other across the different DS comparisons, a fitting format for further analysis. In the manual approach we only focused on the molecules below a *p*-value of 0.05 for all the DS comparisons included independently of their ratio-metric value. However, when calculating the FS index score, since it takes care of the filtering, there is no need of applying cut-off at this step. In case of a statistically poor dataset, where adjusted *p*-values were greater than 0.05, the unadjusted values should be used.

2.3 Grouping and Biomarker Discovery

Dimensionality reduction facilitates analysis and visualization of high-throughput data. This methodology relies on principal component analysis (PCA) to cluster and interpret large-scale DS comparisons. This step represents the 2nd statistical evaluation within this approach. PCA plots allow the identification of outliers, but most importantly it provides a confirmation of the DS comparisons that group together and can be further integrated to perform further analyses. The latter will reduce bias and act as batch effects removal. In order to avoid gaps in the data-matrix and misleading clustering, this analysis should only include and compare the expression-level differences of the molecules analyzed and shared among all DS comparisons.

The 3rd statistical evaluation is founded on pattern matching and centroid clustering to obtain the biomarker list. In the manual approach, DS statistical pattern recognition is based on correlation analysis to generate the biomarker list. Here, unique thresholds (TH) are applied to each DS comparison (THs might vary across DS) depending on the number of their deregulated molecules (where more than 10% of deregulated molecules can give a hint of something off being compared within a DS). Only molecules with significant *p*-values and log 2 fold-change (FC) (above 1 or 2 depending on the DS comparison) are included. A master molecule list (MML) is created containing all the significant molecules reported within THs per DS comparison without repetitions. The MML is used to merge all DS as described above to perform cross-correlation analysis and obtain the biomarker list.

To calculate an accurate frequency distribution manually per molecule, the trend in regulation is determined, taking into account the total count per molecule (TPM), to avoid bias. The regulation trend is described as "Up" or "Down"; molecules reported equally "Up" or "Down" regulated (i.e., 50% up and 50% down regulated) among the DS comparisons are removed. TPM represents the number of times a molecule is analyzed across all the DS comparisons included in the data-matrix. The latter highlights the point that every platform might have different molecular depth, thus, if a molecule is not analyzed in one platform it doesn't mean it is not significant. Consequently, a biomarker list is created and can already be validated. This list is the core of the next functional analysis.

Regarding the FS index, it was developed to generate the former biomarker list from all values across all DS comparisons, without the need of applying individual TH (see below). The FS index is simply based on the $\log_2$FC trend per molecule and whether these are significant or not. The formula is as follows:

$$FS = \frac{\left|\sum(up) - \sum(down)\right|}{\sum(up) + \sum(down)} \times \frac{\sum(significant)}{\sum(all)}$$

From all the DS comparisons, the absolute value of the sum of DS comparisons with up-regulated values ($\log_2$FC > 1) is subtracted from the number of DS comparisons with down-regulation ($\log_2$FC < -1) and divided by the sum of the number of times a molecule was up and down regulated ($\log_2$FC > 1 and <−1). This value is then multiplied by the value obtained from the sum of the number of DS comparisons that have significant

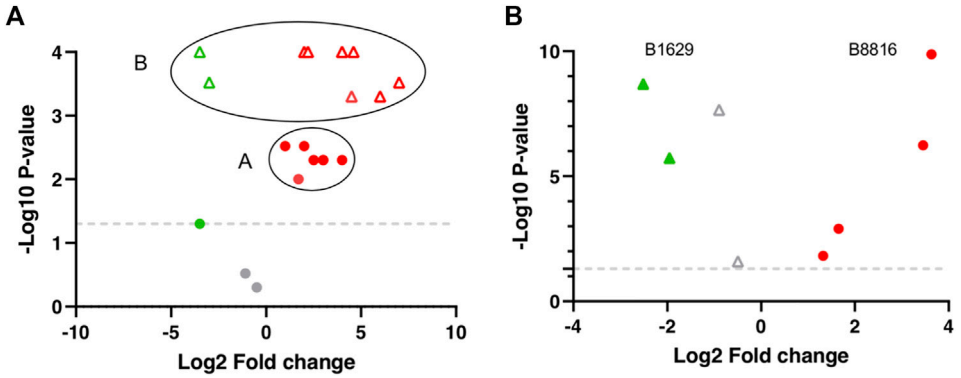

FIGURE 2 | Theoretical and Real centroiding clustering example visualized in a Volcano plot. **(A)** A and B represent 2 different molecules. Red and gray circles represent molecule A distribution from the different dataset (DS) comparisons. The majority of molecule A values cluster regarding regulation, log2FC, and *p*-values (Red circles). Gray circles represent molecule A with non-significant *p*-values. Green and Red triangles represent molecule B distribution. All molecule B values are significant and cluster regarding regulation, log2FC, and *p*-values (Red triangles) but two DS comparisons (Green triangles). **(B)** B1629 and B8816 are real molecules within the DS matrix and represents an example of the distribution of 2 molecules from the biomarker list obtained through the FS index.

values (adjusted *p*-value <0.05) divided by all the DS comparisons that have the specific molecule in question analyzed.

Additionally, the FS index includes an adjusted *p*-value and log2FC calculation. Here, the basis of the approach is centroiding clustering and it applies to all molecules individually. Molecules with *p*-values and FCs for all the included DS can have significant or non-significant values. To visualize the distribution of the data, molecule sets −log (10) of the *p*-values and the log (2) of the FCs can be plotted to get a Volcano plot (**Figure 2**). Here, centroiding clustering brings a logical solution to fuse the data, with the center being the optimal geometric location that minimizes the distance to all datapoints. In an ideal situation the graph (**Figure 2A**) would have a small group of datapoints close to each other for a given molecule (similar *p*-values and similar FCs). The center-value will then provide a new *p*-value and FC value. To obtain the adjusted *p*-value its easier than the adjusted $\log_2$FC value, since it does not have directionality; in this case an arithmetic mean is calculated from the −log10 *p*-values to obtain the adjusted *p*-value, since the non-transformed *p*-values are too small and here a TH does not apply, therefore significant and non-significant *p*-values are included and averaging these can introduce bias. Regarding the adjusted log2FC, a common mean calculation might give a biased result due to the possible large values with opposing trends that the different DS comparisons can have (**Figure 2B**). There are still several options one can follow, such as geometric weighted means, where a specific value is added to each log2FC, e.g., sample size; trimmed means where extreme values (outliers) are left out, and the mean is calculated from the values that remain (Lawson et al., 2012; Miao and Jiang 2014; Li X. et al., 2015). However, in this example and regarding the FS index calculation, only where >50% of the molecules follow the same trend (up or down regulation) with values above or below 0 molecule value were averaged, the rest of the molecules were excluded.

These data, plus the calculated FS index value, upgrade the score system and allow us to rank the molecules from the most to the least significant based on: total significance, trend (up/down-regulation), number of times a molecule was analysed and present an either up or down-regulation, adjusted *p*-value and adjusted log2FC, which sums up into the FS index score. The higher the score, the more significant a molecule is.

2.4 Data Integration and Functional Analysis

Despite the biomarker list potential to lead to new insights about the disease in question, the contextualization of these molecules can be even more informative. In this section of the approach, the integration of the biomarker list through enrichment analysis is performed. Functional analysis through Cytoscape plug-ins ClueGO/CluePedia performs semantic clustering by assigning gene ontologies and/or pathway terms (KEGG, Wikipathways, Reactome) to the biomarker list, integrates them into functional networks and ties-in the molecules associated with each of the terms on the networks generated (Shannon et al., 2003; Bindea et al., 2009; Bindea et al., 2013).

This analysis will highlight the main deregulated processes that will be the center for further analyses. The biomarker list contains molecules with different frequencies and FS index scores, and results from the merging of the different DS comparisons. Several TH are applied to the biomarker list to determine the main deregulated pathways within it. The THs go from the most to the least significant and frequent molecules within the biomarker list. The TH end cut-off goes to a level where the processes identified through the analysis of the most frequent molecules are not lost but enriched and interconnected by the molecules from the different THs applied.

In order to visualize and underlie the main processes previously identified, pathway mapping is performed. It helps unravel the regulation and involvement of the molecules by placing them within their described position in the pathway of interest. This allows the accurate identification of deregulated processes by showing specific trends through the molecular

interplay, hence the identification of key players involved in the pathophysiology of a disease. Pathvisio is a pathway-map editor of described and pre-assembled pathway maps (KEGG, Wikipathways, Reactome), it is a fine tool for integrative analysis since it handles gene, protein, and metabolite data and allows cross-mapping and integration through Bridgedb (van Iersel et al., 2010; Kutmon et al., 2015). The pre-assembled pathway maps will function as the sketch to base on to get novel pathway maps reflecting their regulation in the setting of interest. The pathways to be filled-in, edited, and integrated are the ones identified by ClueGO/CluePedia analysis. First, the molecules from the biomarker list are mapped into their specific position within these pathways to help keep the focus. Afterwards, irrespective of their logFC, all the molecules from the first data merging with a significant *p*-value are mapped as well in order to fill the gaps within the pathway map of interest and be able to identify trends in regulation.

To enrich and complement the processes of interest identified through Pathvisio, interactome analyses are performed on the biomarker list molecules. GeneMANIA (Multiple Association Network Integration Algorithm) generates networks that resemble molecular interactions classified into gene-protein interaction, co-expression, and localization, shared protein domains, and pathways (Warde-Farley et al., 2010). It provides the connections of the biomarker list molecules and predicts molecules associated with the input.

In addition, disease analysis to explore the former outcomes and their accurate association with the specific disease of interest is performed. Through DisGeNET, reported gene-disease associations from the biomarker list are identified (Piñero et al., 2015; Piñero et al., 2020). This step functions as validation by detecting the genes that have already been reported in the disease under study. DisGeNET also provides genes with gene-disease-association (GDA) scores, and the ones with a higher score can be used to enrich the pathway model by identifying their associations with the biomarker list molecules and thus, provide an extra focus on the pathways where these molecules play a role.

De novo pathway contextualization is produced by the integration of all the different results previously obtained. Since transcriptomic studies populate the databases and literature, these are the backbone of the methodology. When analyzing different molecular types, each OMIC layer validate and reinforce the focus on the deregulated mechanisms identified through transcriptomics analysis. miRNAs (miRs) and metabolites do not align with gene/proteins but can also be integrated into the developed model. In this case, cross-mapping can be carried out through "mode of action" by using either their targeted genes as a substitute ID (miR) or how they are produced (metabolites) using the associated enzyme to tie them into other OMICS data. By using the common unifier, it allows their correlation and mapping into the *de novo* pathway model described. CluePedia and MetaboAnalyst web-tools can serve this purpose by enriching miRs and metabolites respectively (Pang et al., 2020). In this example, only transcriptomics data is included.

3 RESULTS

3.1 DLBCL Dataets Comparisons Correlation and Grouping

A total of 6 DLBCL gene expression profiles from human samples were identified through GEO, correlated, and meta-analyzed through the HHmeta method. From 6 GEO DS we ended up with 10 comparisons: GSE9327–1. DLBCL vs Healthy tissue; GSE32018–2. DLBCL vs. Healthy tissue; GSE56315–3. Plasmablast DLBCL vs. Plasmablast Healthy B-cell, 4. Centroblast DLBCL vs. Centroblast Healthy B-cell, 5. Centrocyte DLBCL vs. Centrocyte Healthy B-cell, 6. DLBCL vs. Healthy B-cells; GSE12195–7. DLBCL vs. Healthy B-cells; GSE2350–8. DLBCL vs. Healthy B-cells; 9. DLBCL CD19 B-cells vs. Healthy B-cells; GSE12453–10. DLBCL B-cells vs. Healthy B-cells. All molecules from the DS comparisons were mapped and indexed to PADB unifier ID. Only molecules/probe-sets with the most significant *p*-values were kept among the repeats found within each DS comparison. For the manual approach, DS comparisons were filtered by *p*-value (<0.05), regardless of their logFC value, and merged. When following the HHmeta method, no filtering is needed at this stage.

Dimensionality reduction through PCA clustered 2 groups and identified a potential outlier (**Figure 3A**). DS 1, 2, 8, 9, and 10 (Group 1) and DS 3, 4, 5, and 6 (Group 2) were clustered together. Group 1 is composed of comparisons across healthy and tumoral tissue (DS 1 and 2), as well as healthy and tumoral B-cells (DS 9 and 10), however DS 8 compares both healthy B cells vs DLBCL tumoral tissue and also groups within this cluster. Group 2 contains 1 solely GEO DS (GSE56315) that is composed mainly by comparisons among specific tumoral tissue DLBCL subtypes and their matching healthy B-cell type, as well as the comparison of all of them together plus some unclassified DLBCLs and the complete population of healthy B-cells. Group 2 samples belong to patients under either CHOP or R-CHOP therapy. The highlighted outlier, DS 7 is composed of DLBCL tissue samples and healthy tonsillar germinal center, naive and memory B cells.

3.2 Biomarker List and ClueGO/CluePedia Functional Analysis

Once grouped, the following analysis focused only on group 1. For the manual procedure, *p*-value cut-off (<0.05) from the previous filtering was kept and log2FC cut-offs (>1 and <−1) were applied to each DS comparison. When applying the HHmeta method FS index formula (see above), no threshold is needed. Data merging allowed the comparison and consolidation of molecular regulation based on their frequency distribution. The resultant biomarker list from the manual procedure contains a total of 3,241 significant molecules, and from the FS index calculation, a total of 1,638 significant molecule (**Supplementary Table S1**). **Table 2** represents the top up and down-regulated molecules from both biomarker lists group 1.

Through DisGENET one can already search for validation of genes associated with the condition in question within the

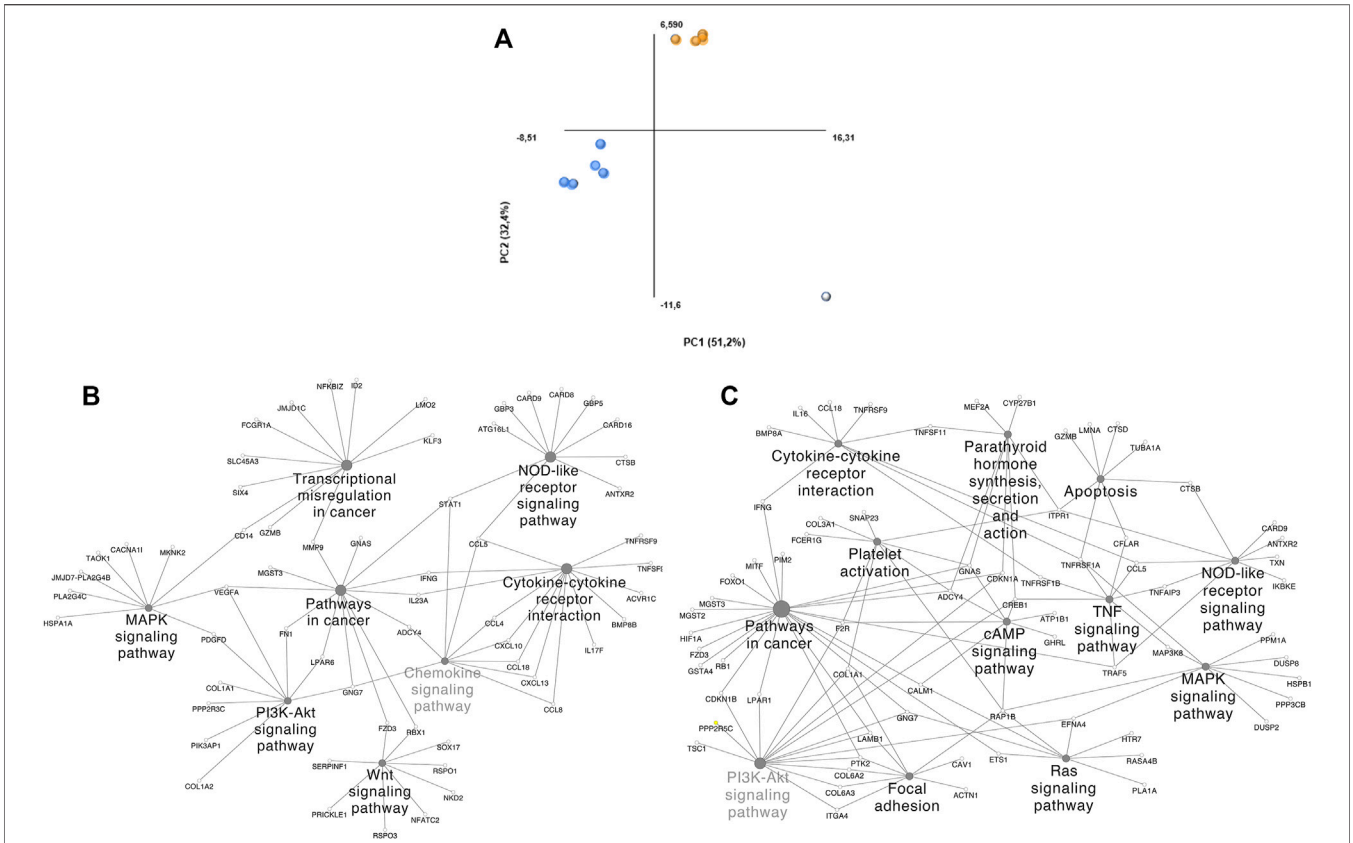

FIGURE 3 | PCA plot grouping and ClueGO/CluePedia focus of our DLBCL cohorts. **(A)** Blue dots: Group 1; Orange dots: Group 2; White dots: Outliers. **(B)** ClueGO/CluePedia network created from the Manual approach biomarker list (66% threshold). **(C)** ClueGO/CluePedia network created from the FS index Biomarker list (0.75 Absolute FS index threshold).

TABLE 2 | Top deregulated molecules obtained with the Manual approach and FS Index calculation.

ID		Manual approach	HHmeta method		
CluSO ID	**Gene name**	**Final Regulation**	**Adj. P.V. Mean**	**Log2FC Mean**	**FS Index Calculation**
B2Q85	ITGA9	100	9.130E-23	4.29	1
B2O29	BIRC3	100	2.480E-05	2.19	1
BO058	HLA-DRB1	100	3.320E-03	2.04	1
BO135	BCL6	100	1.210E-04	2.01	1
B8773	LCE2D	−100	2.760E-03	−2.24	1
B9009	LPP	−100	4.687E-05	−2.25	1
BF875	SYCE1L	−100	1.270E-05	−2.29	1
B1137	ATP10D	−100	8.127E-08	−2.30	1
BL316	DNM1DN8-2	−100	3.360E-04	−2.40	1
BH305	TSPYL5	−100	4.326E-07	−2.49	1
B5415	FAM208B	−100	7.820E-07	−2.55	1
B7596	IGF2	−100	3.810E-08	−2.59	1
BO237	RET	−100	1.400E-17	−2.70	1
B8780	LCE5A	−100	1.660E-04	−2.87	1
B8612	KRTAP5-3	−100	2.240E-07	−3.00	1
B6621	GPR150	−100	3.360E-07	−3.01	1
B2W20	YES1	−100	2.380E-08	−3.45	1
B7559	IFITM5	−100	3.360E-07	−3.46	1

TABLE 3 | Top genes associated with DLBCL through DisGENET.

Gene	GDA Score	Association Type	Number of PMIDs
BCL2	0.4	Biomarker Altered Expression Genetic Variation	222
FBXO11	0.32	Biomarker Genetic Variation Causal Mutation	2
IRF8	0.32	Biomarker Altered Expression Causal Mutation	2
BCL6	0.1	Biomarker Altered Expression Genetic Variation	224
BIRC3	0.08	Biomarker Genetic Variation	8
HDAC9	0.07	Biomarker Altered Expression	7
ZC3H12D	0.05	Biomarker	5
LIG4	0.04	Biomarker Post-translational modification	4
HLA-DRB1	0.03	Biomarker Genetic Variation	3
PSIP1	0.02	Biomarker	2

biomarker list generated or target genes obtained from the functional analysis. In this example biomarker list genes with a FS index equal to 1 (439 genes) were input into DisGENET. As shown in **Table 3**, some have already been reported as related to DLBCL, either as e.g., biomarker, altered expression or genetic variants. DLBCL is a highly heterogeneous disease, thus it is important to bear in mind that in this example we didn't segregate DLBCL by type (e.g., GCB, ABC), mainly because it was not specified in every dataset included. Therefore, the outcome of this analysis would mirror the core shared mechanisms among the DLBCLs analyzed in each of the different studies included in the meta-analysis of group 1. Also, within the genes found to be associated with DLBCL from the biomarker list, these might have been significantly deregulated in only one dataset because it was the only one analyzing this molecule, however this doesn't rule out its importance since these can still interact with the main pathways further identified. For instance, genes involved in NF-kappa B pathway and TNF signaling, such as overexpression of BCL2 regulator of mitochondrial apoptotic pathway (Tsuyama et al., 2017), BCL6 proto-oncogene, essential for GC development and FBXO11 a tumor-suppressor gene that stabilizes BCL6, have already been related with DLBCL accelerated development and poor prognosis (Saito et al., 2007; Duan et al., 2012; Zhang et al., 2015). These, plus the rest of molecules covered by DisGENET provide validation of the molecular list we relied on for further functional analysis, and the meta-analysis itself.

Besides the already described molecules in the DLBCL setting, the biomarker list contains molecules that have not been related with the disease yet. As an example, one of the main deregulated genes with a higher frequency distribution score is TSPYL5, which was found with a downregulation trend in 4 out of the 5 DS comparisons included in group 1. TSPYL5 has been attributed a tumor-suppressive function, and its hypermethylation has been previously linked with several cancers (Kim et al., 2010; Fan et al., 2020; Huang et al., 2020). TSPYL5 suppression has been associated with PTEN overexpression and AKT pathway inhibition (Vachani et al., 2007; Jung et al., 2008; Kim et al., 2010; Fan et al., 2020). Interestingly, TSPYL5 inhibition has been attributed to overexpression of miR-483–5p and miR-629. In prostate cancer it's been recently proven that miR-483–5p antagonization through the long non-coding RNA LINC00908 lead to an upregulation of TSPYL5, inhibiting prostate cancer progression (Fan et al., 2020). miR-629 overexpression has also shown the ability to promote proliferation, migration and invasion in ovarian cancer by directly inhibiting TSPYL5 (Shao et al., 2017). Although a specific role in DLBCL has not being described yet, these results open a potential novel regulation of carcinogenesis in this setting. ATP10D is also within the genes with a higher FS index score. Although its association with DLBCL hasn't been described yet, its downregulation has been significantly correlated with poor non-small cell lung cancer survival (Fusco et al., 2018). It belongs to a subfamily of P-type ATPases that play a role in phospholipids translocation, and its being specially associated with sphingolipids and ceramid plasma levels (Hicks et al., 2009). Sphingosine-1-phosphate (S1P) sphingolipids are considered signaling molecules involved in activation of carcinogenesis pathways and have been previously linked to increase lung cancer risk (Furuya et al., 2011; Alberg et al., 2013). These are interesting hypothesis that haven't been explored yet that by following our unbiased method could be highlighted. Examples like these can already be validated, adding to the main DLBCL pathway mechanisms.

3.3 ClueGO/CluePedia Functional Analysis

To distinguish the association amongst the biomarker list molecules and determine their shared pathways and processes, ClueGO/CluePedia analyses were performed (**Figures 3B–C**). The main processes showing interconnectivity between the molecules from the biomarker list of both, the manual approach and FS index score were MAPK, PI3K, TNF, Ras and B-cell signaling pathways, cytokine-cytokine receptor interaction and chemokine signaling pathways, among others. These pathways show high interconnectivity and potential involvement of MMP9, STAT1, TNFRSF1A, NFKBIA, EFNA4, CCL5, RAP1B. Networks in **Figures 3B,C** are similar, even though the HHmeta method does not follow thresholds and has an extra layer of significance raking through the FS index score calculation. However, since the most significant molecules were shared among the biomarkerlists generated through the manual approach and the HHmeta method, the main deregulated processes remain.

All the main pathways highlighted through this analysis are somehow related with pro-survival signaling, and have been previously associated with DLBCL pathology, as well as with the heavy involvement and crucial role of the tumor microenvironment. Pro-survival effects via PI3K-AKT

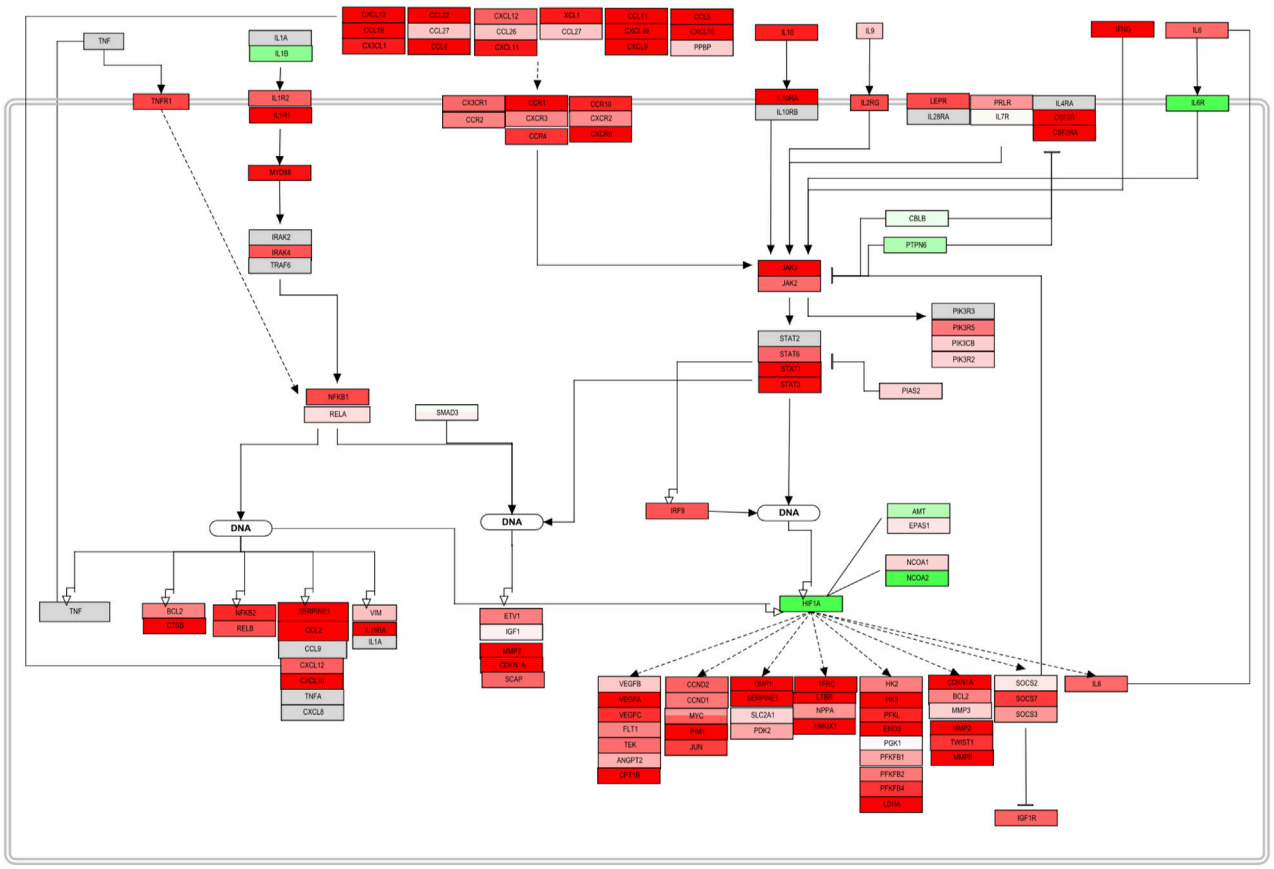

FIGURE 4 | PathVisio edited pathway of the obtained biomarker list. NFKB and JAK/STAT section of the complete pathway map from **Supplementary Figure S1**. This section of a pathway contextualizes and represents the up-regulation trend of the molecules included in the map. Molecules with an adjusted *p*-value <0.05 from the FS index score calculation were included.

signaling pathway, Ras signaling pathway (Eric Davis et al., 2001; Davis et al., 2010; Miao et al., 2019), partly B-cell receptor signaling pathway and cytokine induction have been long correlated with DLBCL. TNF signaling pathway is known to be indispensable for survival of transformed B-cells. TNF-signaling pathway regulates NF-kappa B pathway and MAPK signaling pathway, which was also determined as significant in DLBCL (Webster and Vucic 2020).

3.4 Pathvisio Pathway Editing and Complementation

The main pathways determined by ClueGo/CluePedia are now involved in the following iteration step, which entails pathway overrepresentation analysis and visualization in Pathvisio. Pathways identified through ClueGO were displayed within the significant pathways determined by Pathvisio. In order to determine consistencies, inconsistencies, interconnected events, and fill the gaps within the signaling pathways visualized, the complete list of molecules with a significant *p*-value is used as an input. In this case, a total of 5,146 molecules from the FS index were analyzed. For instance, the total amount of molecules obtained from the FS index might differ from the ones of the manual approach. This because for the adjusted log2FC calculation, the FS index only takes molecules determined as up or down regulated for averaging (see **section 2.3**), leaving out some molecules with mixed values. The manual approach is more bias, where the arithmetic mean is calculated from all molecules with a significant *p*-value, without prior filtering.

Once mapped, inconsistencies in the trend from sections or complete signaling events were removed from the original pathway maps. The complete pathway can be visualized in the **Supplementary Figure S1**. The original pathway maps from WikiPathways were redesigned to accurately contextualize the role and interplay of the molecules in DLBCL B-cells. Here, only a section of the assembled pathway is shown in **Figure 4**. From this signaling map, a prominent up-regulation of most of the elements involved can be seen. Chemokines, cytokines, and interleukin signaling demonstrate their involvement in the NF-KB and JAK-STAT pathway activation and therefore cell survival and proliferation. Several factors either display an opposite regulation or are absent within the biomarker list generated. However, the consistency along the pathway outline reflects an involvement of these processes in this B-cell malignancy.

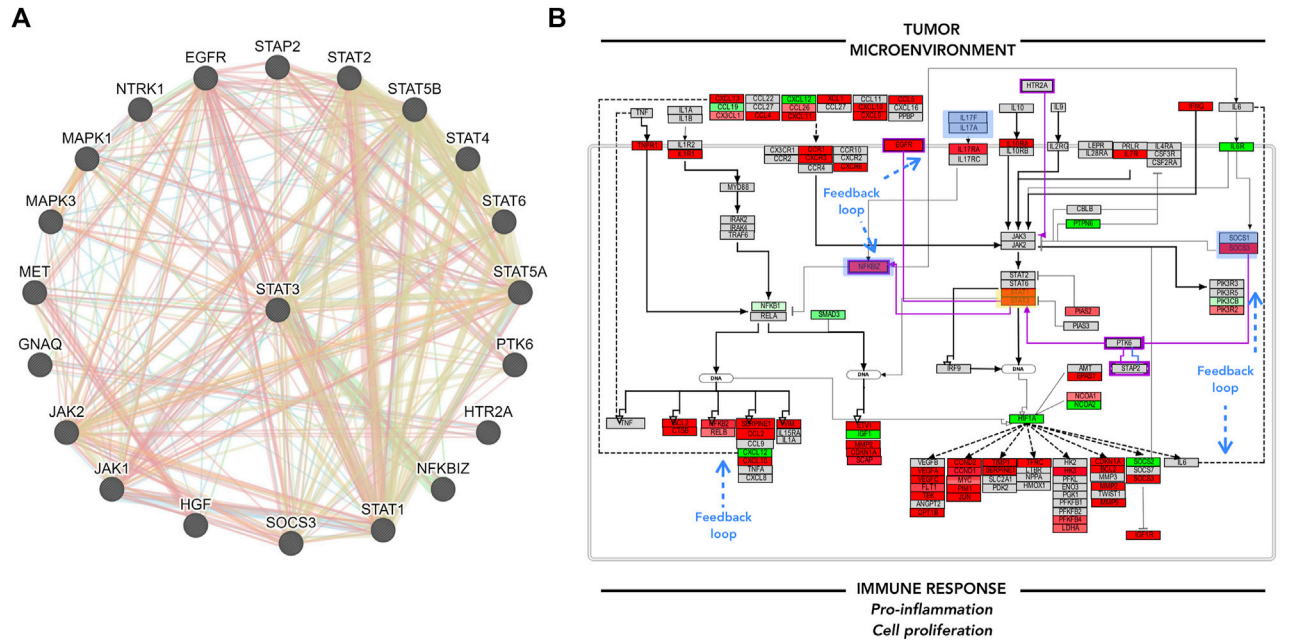

FIGURE 5 | GeneMANIA focus and Pathvisio *de novo* pathway contextualization. **(A)** GeneMANIA results for STAT3; interactor molecules can either represent physical interaction (red), co-expression (purple), genetic interaction (green), shared protein domain (yellow). **(B)** Pathvisio NFKB and JAK/STAT signaling pathway section with added elements from GeneMANIA highlighted in purple; STAT3 analyzed gene highlighted in yellow box, blue boxes represent feedback loops. analyzed gene highlighted in yellow box, blue boxes represent feedback loops.

Feedback loops such as the one established by IL6, as well as regulators, such as PIAS, SOCS, and PTPN6 can already be spotted and represent potential hits to focus on further (**Figure 5B**).

GeneMANIA analysis can enrich the pathway maps from PathVisio. Several angles can be considered to add to the main hypothesis, for example, the top molecules from the biomarker list, the associated and clustered molecules obtained from ClueGO/CluePedia analysis, or the molecules of interest within the edited PathVisio pathway. Here, STAT3 was taken as an example to complement the pathway of interest, due to its downstream effects and the potential regulation of its activity. **Figure 5**. A. shows GeneMANIA results. Several molecules that were not included in the pathway map make an appearance. The connected molecules can either represent genomic interaction, shared protein domain, shared pathway, co-expression, co-localization, among other interactions. The additional interactors of STAT3 identified through GeneMANIA, such as NFKBIZ, EGFR, PTK6, and STAP2, complement the previous pathway, and the trend in consistency remains (**Figure 5B**). Here, one can hypothesize that the tumor microenvironment, such as T-cells, promote the chemokine/cytokine signaling in B-cells and lead to pro-inflammation, cellular proliferation and maintenance to some extent through JAK/STAT and NFKB signaling pathway constant activation. The interconnected and functionally correlated genes identified through ClueGO/CluePedia analysis, and some others integrated into the final edited pathway, have shown to be already associated with DLBCL.

4 DISCUSSION

The decreased costs of high-throughput technologies have made the exploratory studies of complex biological traits, such as cancer, possible. Integrative omics approaches have been under the spotlight due to their potential to elucidate novel pathophysiological insights that better capture the complexity of molecular systems in a trait (Argelaguet et al., 2018; Kim et al., 2018; Zhou et al., 2019). Despite the increase in studies performing this type of analysis, efforts are still needed to better analyse and decipher the origin of complex diseases, for better diagnostics and discovering potential therapeutic targets, reviewed in (Yan et al., 2017; Karczewski and Snyder 2018). As a common characteristic, integrative methodologies rely on the identification of shared features across different large-scale datasets to further perform functional analysis. Nevertheless, one of the main elusive challenges that remains is the contextualization of the deregulated molecules; particularly in cancer where the high variability and intricacy of biomolecules involved can overwhelm meaningful readouts. In this setting, it is complex to identify commonalities among the systems altered by only looking at molecular signatures or protein-protein interactions, even within samples from the same cancer type. Thus, even though novel insights regarding potential correlations have been depicted across multi-omics studies (Zhang et al., 2013; Li J. et al., 2015; Mertins et al., 2016), the contextualization of how a molecule might influence or affect a system is still lacking. Our proposed methodology focuses not only on the identification of shared features, but also on their contextualization through

pathway mapping. Approaches such as IPA, also focuses on the contextualization of features into pathway maps, however the lack of identifier curation, track and maintenance can result in poor reproducibility. In high-throughput studies the sample size is also an issue that might affect reproducibility and specificity. Usually, sample size increment correlates with higher reproducibility; it is equally responsible for an increase in false positives (Maleki et al., 2019). Thus, the fusion of integrative approaches, meta-analysis, associative data and enrichment methodologies gives an opportunity to boost the understanding, correlation and contextualization of potential molecules of interest affected in a disease. Moreover, one of the main hallmarks of our methodology is the enhancement of the statistical power of the biomarkers identified not only through the integration of high-throughput studies but also small-scale studies, which provides the focus on the pathway maps further described.

The majority of the available big data approaches rely on computational tools and therefore, the need of certain background to be able to perform these analyses. However, the HHmeta method provides a platform to not only perform an integrative meta-analysis, but also the opportunity for researchers lacking a solid background in bioinformatics to be able to perform an unbiased and straight-forward, but robust meta-analysis on pre-processed big data, to reach a logical and contextualized overview of the molecular interplay of a list of significant molecules related to an specific research question. In the example presented above, besides the identification of a deregulated and correlated set of molecules through out the analysis of different studies (**Table 2**), this methodology allowed their contextualization to identify potential processes and mechanisms involved in the disease (**Figures 3**, **4**), and clarified targets influencing cell growth, survival and metastasis.

An interesting aspect that is commonly under-rated–but can influence downstream analysis and affect its replication–is the wide range of identifiers. Their constant update, reuse, un-usage, and the lack of unified efforts to both keep track and make mapping between different platforms available for the scientific community. Moreover, in order to merge datasets from different platforms and sources, the harmonization of identifiers is crucial. Thus, one of the solid basis and uniqueness of this methodology is its reliability on the PADB unifier database (see Methods section). Efforts have been previously made by others to address this issue (Gaj et al., 2007; Klimke et al., 2011), through BLASTx approaches (e.g., TargetIdentifier), linking annotations from different databases (e.g., DAVID) and trying to provide as much information as possible about IDs (Gaj et al., 2007), however data curation and constant update its still lacking. It has been noticed in pathway mapping that a great proportion of arrays become useless, mainly because there is no track of older IDs. Herein, PADB adds an extra quality-check to be able to rely fully on the available annotations and support the replication process. It enriches the downstream biological pathway map interpretation by retaining old identifiers for those molecules that currently have no annotation. PADB also allows cross-linking through species by its ortholog IDs (OMAP), enabling the identification of mechanisms that might be conserved across species through the downstream analysis.

Conventional meta-analyses apply several strategies to merge statistical measurements (i.e., *p*-value), and this is one of the main differences highlighted in the methodology presented here. Methods such as Fisher's, Stouffer's *Z*-test and Rank product are examples of popular statistical approaches to follow when performing meta-analyses to combine *p*-values of different studies, and their use depends on the meta-analysis goal (Zaykin et al., 2007; Hong and Breitling 2008). The former is based on testing the probability that different null-hypotheses, when combined, are statistically significant (Fisher 1992). However, here the proposed methodology relies on already statistically significant data for the manual approach; the FS index calculation (see **section 2.3**) relies on the number of times a molecule is significant, and the adjusted significance of a molecule is only one layer of ranking to consider. Therefore, in this setting Fisher's method would not be the one of choice. In this methodology, the more a molecule is significant across studies the more likely it is that it is significant overall, regardless of the actual *p*-value, same with log2FCs, threshold values are used to set those boundaries. The significance of the molecules identified through this method is then corroborated by pathway mapping and other further analyses. Nevertheless, the HHmeta method and Fisher's are similar in the principle of getting a new *p*-value (in our case also log2FC, plus the FS index) through the fusion of all the studies included. The main difference among conventional meta-analyses and our proposed methodology is that, by keeping the studies intact regarding cases and controls, and correlating the DS comparisons *p*-values and log2FCs, this methodology adds a layer of confidence regarding the comparisons made, allowing a primary correlation and clustering of studies through PCA plots.

PCA plot analysis have been used for the purpose of modelling the relationship between samples, to detect group differences and identify outliers and batch effects within a single high-throughput study (Ringnér 2008; Conesa et al., 2016; Merino et al., 2016; Todorov et al., 2018). Furthermore, there have been other integrative methodologies that have generalized PCA rationale to identify commonalities across different omics studies (Kim et al., 2017; Argelaguet et al., 2018). In contrast, here we apply PCA plot analysis to individual omics data types. Through this analysis, we were able to identify different groups and outliers from the initial high-throughput studies included in the analysis. This quality-check gives the opportunity to identify commonalities even amongst different samples, such as complete tumoral tissue and B-cells, by only including the molecules analysed in all the studies. It allows us to subtract commonalities across diverse studies, provided that the research question is well established. Even though the meta-analysis performed through this methodology is based on similar data, and therefore group of studies, it allows the comparison and identification of the relation between groups of different conditions (e.g., DLBCL vs. Healthy and DLBCL treated vs. DLBCL non-treated) opening new opportunities to identify specific responses and common enhanced pathways deregulated by the disease itself. Thus, to perform standard meta-analysis will be inadequate.

The major liability of the HHmeta method is that it is based on publicly available data. Thus, it is possible that the specific research question one wants to address hasn't been covered by many research groups. The less high-throughput data sets available for a certain topic of interest, the less statistical power the data analysis would have. Moreover, if the available data is heterogenous, for example due to differences in the biology of the samples (treatments, stages of a disease, subtypes or sample sources), it makes the correlation even more complex, and the main question to address would need to change into a more general one, where commonalities can be depicted. Another weakness of the whole procedure is that the contextualization of the biomarker list relies on pathways previously described, so there will be gaps, molecules that do not map and unsolved questions. Despite these flaws, this method takes advantage of previous knowledge and uses it in the context of the specific topic of interest.

The substantial amount of data generated throughout the years represent a tool that can be somehow overlooked by the scientific community. For instance, good scientific practice can be enhanced by the screening, review and statistical analysis of previous studies performed in the field of interest to identify the gaps, commonalities and generating a better understanding regarding the behaviour of a system of interest, by feeding a potential model with what has already been proven and enhancing the generation of novel hypothesis to address by the inclusion of high-throughput data. The essence of the proposed methodology is the merging of independent statistical tests in an unbiased way, into a single test. It embraces the availability and basis of statistical analyses used in the big data field and utilizes their outcome to add to the statistical power of the data, resulting in a novel analysis approach. What sets the HHmeta method apart from the already available approaches are the basis of the data considered for merging, thresholding and its subsequent fusion and scoring system. Here, the manual and FS index approach are presented to highlight the main differences of what has been done before with the same ground basis as the FS index approach (FS index) which relies on a formula and adjusted values to produce similar results.

5 CONCLUSION

Studies in basic science are commonly hypothesis-driven and usually small-sampled. Likewise, and despite their exploratory nature, high-throughput studies tend to be biased to the resulting top deregulated genes. Therefore, novel findings require further validation, and here is where meta-analysis comes in handy. Even though the experimental design of different studies in essence is unique, meta-analysis methodologies have provided the opportunity to integrate the results of diverse and multiple studies addressing the same question, to enhance the statistical power of the results and therefore, the chances of finding true positives. In contrast to the meta-analysis methodologies already implemented in the big data-field, the methodology presented in this manuscript provides a simple, novel, unbiased, integrative and logical approach to not only meta-analyze single omics studies, but to integrate small and big data sets, as well as different omics studies. It includes quality checks to avoid batch effects, relies on a powerful cross-indexing unifier database and goes a step further by including associative data to aid for the identification and understanding of novel pathways and molecules involved in a specific disease. All in all, the current methodology provides novel hypotheses to further validate and a broader view of the system of interest, enhancing the outcomes generated through conventional meta-analysis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

HH conceptualized and designed this study. KG performed the data collection and analysis. KG wrote the draft manuscript. RC and HH participated in manuscript revision. All authors contributed to the article and approved the submitted version.

FUNDING

KG is supported by CONACYT Mexico scholarship (No. 2019-000021-01EXTF-00542). HH is supported by a grant from Highlands & Islands Enterprise. RC is supported by the Biotechnology and Biological Sciences Research Council (BBSRC; BB/N017773/2), the Swiss National Science Foundation (SNSF; CRSK-3_190550), the Rosetrees Trust Fund (M713), and the University of Zürich Research Priority Program (URPP–Translational Cancer Research).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.828786/full#supplementary-material

REFERENCES

Alberg, A. J., Armeson, K., Pierce, J. S., Bielawski, J., Bielawska, A., Visvanathan, K., et al. (2013). Plasma Sphingolipids and Lung Cancer: A Population-Based, Nested Case-Control Study. *Cancer Epidemiol. Biomarkers Prev.* 22, 1374–1382. doi:10.1158/1055-9965.EPI-12-1424

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-Omics Factor Analysis-A Framework for Unsupervised Integration of Multi-omics Data Sets. *Mol. Syst. Biol.* 14, e8124. doi:10.15252/msb.20178124

Arneson, D., Bhattacharya, A., Shu, L., Mäkinen, V.-P., and Yang, X. (2016). Mergeomics: A Web Server for Identifying Pathological Pathways, Networks, and Key Regulators via Multidimensional Data Integration. *BMC Genomics* 17, 722. doi:10.1186/s12864-016-3057-8

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., et al. (2019). ArrayExpress Update - from Bulk to Single-Cell Expression Data. *Nucleic Acids Res.* 47, D711–D715. doi:10.1093/nar/gky964

Auffray, C., Adcock, I. M., Chung, K. F., Djukanovic, R., Pison, C., and Sterk, P. J. (2010). An Integrative Systems Biology Approach to Understanding Pulmonary Diseases. *Chest* 137, 1410–1416. doi:10.1378/chest.09-1850

Badr, M. T., and Häcker, G. (2019). Gene Expression Profiling Meta-Analysis Reveals Novel Gene Signatures and Pathways Shared between Tuberculosis and Rheumatoid Arthritis. *PLoS One* 14, e0213470. doi:10.1371/journal.pone.0213470

Bell, R., Barraclough, R., and Vasieva, O. (2017). Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers. *Cmm* 17, 200. doi:10.2174/1566524017666170807144946

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodological)* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bindea, G., Galon, J., and Mlecnik, B. (2013). CluePedia Cytoscape Plugin: Pathway Insights Using Integrated Experimental and In Silico Data. *Bioinformatics* 29, 661–663. doi:10.1093/bioinformatics/btt019

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: A Cytoscape Plug-In to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks. *Bioinformatics* 25, 1091–1093. doi:10.1093/bioinformatics/btp101

Boeing, S., Williamson, L., Encheva, V., Gori, I., Saunders, R. E., Instrell, R., et al. (2016). Multiomic Analysis of the UV-Induced DNA Damage Response. *Cel Rep.* 15, 1597–1610. doi:10.1016/j.celrep.2016.04.047

Care, M. A., Westhead, D. R., and Tooze, R. M. (2015). Gene Expression Meta-Analysis Reveals Immune Response Convergence on the IFNγ-STAT1-IRF1 axis and Adaptive Immune Resistance Mechanisms in Lymphoma. *Genome Med.* 7, 96. doi:10.1186/s13073-015-0218-3

Carroll, A. J., Badger, M. R., and Harvey Millar, A. (2010). The MetabolomeExpress Project: Enabling Web-Based Processing, Analysis and Transparent Dissemination of GC/MS Metabolomics Datasets. *BMC Bioinformatics* 11, 1–13. doi:10.1186/1471-2105-11-376

Cervantes-Gracia, K., Chahwan, R., and Husi, H. (2021). Of Incongruous Cancer Genomics and Proteomics Datasets. *Methods Mol. Biol.* 2361, 291–305. doi:10.1007/978-1-0716-1641-3_17

Cervantes-Gracia, K., and Husi, H. (2018). Integrative Analysis of Multiple Sclerosis Using a Systems Biology Approach. *Sci. Rep.* 8, 1–14. doi:10.1038/s41598-018-24032-8

Cervantes-Gracia, K., Gramalla-Schmitz, A., Weischedel, J., and Chahwan, R. (2021). APOBECs Orchestrate Genomic and Epigenomic Editing Across Health and Disease. *Trends Genet.* 37, 1028–1043. doi:10.1016/j.tig.2021.07.003

Cho, H., Kim, H., Na, D., Kim, S. Y., Jo, D., and Lee, D. (2016). Meta-analysis Method for Discovering Reliable Biomarkers by Integrating Statistical and Biological Approaches: An Application to Liver Toxicity. *Biochem. Biophysical Res. Commun.* 471, 274–281. doi:10.1016/j.bbrc.2016.01.082

Clough, E., and Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods Mol. Biol.* 1418, 93–110. doi:10.1007/978-1-4939-3578-9_5

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A Survey of Best Practices for RNA-Seq Data Analysis. *Genome Biol.* 17, 1–19. doi:10.1186/s13059-016-0881-8

Consortium, T. U., Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., et al. (2021). UniProt: the Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi:10.1093/NAR/GKAA1100

Coutant, S., Cabot, C., Lefebvre, A., Léonard, M., Prieur-Gaston, E., Campion, D., et al. (2012). EVA: Exome Variation Analyzer, an Efficient and Versatile Tool for Filtering Strategies in Medical Genomics. *BMC Bioinformatics* 13, 1–12. doi:10.1186/1471-2105-13-S14-S9

Davis, R. E., Brown, K. D., Siebenlist, U., and Staudt, L. M. (2001). Constitutive Nuclear Factor κB Activity Is Required for Survival of Activated B Cell-like Diffuse Large B Cell Lymphoma Cells. *J. Exp. Med.* 194, 1861–1874. doi:10.1084/jem.194.12.1861

Davis, R. E., Ngo, V. N., Lenz, G., Tolar, P., Young, R. M., Romesser, P. B., et al. (2010). Chronic Active B-Cell-Receptor Signalling in Diffuse Large B-Cell Lymphoma. *Nature* 463, 88–92. doi:10.1038/nature08638

Davis, S., and Meltzer, P. S. (2007). GEOquery: A Bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi:10.1093/bioinformatics/btm254

Deutsch, E. W. (2010). The PeptideAtlas Project. *Methods Mol. Biol.* 604, 285–296. doi:10.1007/978-1-60761-444-9_19

Duan, S., Cermak, L., Pagan, J. K., Martinengo, C., di Celle, P. F., Chapuy, B., et al. (2012). FBXO11 Targets BCL6 for Degradation and Is Inactivated in Diffuse Large B-Cell Lymphomas. *Nature* 481, 90–93. doi:10.1038/NATURE10688

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and Bioconductor: A Powerful Link between Biological Databases and Microarray Data Analysis. *Bioinformatics* 21, 3439–3440. doi:10.1093/bioinformatics/bti525

Fan, L., Li, H., and Zhang, Y. (2020). LINC00908 Negatively Regulates microRNA-483-5p to Increase TSPYL5 Expression and Inhibit the Development of Prostate Cancer. *Cancer Cel Int* 20, 10. doi:10.1186/s12935-019-1073-x

Fenyö, D., and Beavis, R. C. (2015). The GPMDB REST Interface. *Bioinformatics* 31, 2056–2058. doi:10.1093/BIOINFORMATICS/BTV107

Fernandes, M., and Husi, H. (2016). Integrative Systems Biology Investigation of Fabry Disease. *Diseases* 4, 35. doi:10.3390/diseases4040035

Fernandes, M., and Husi, H. (2019). *Integrative Systems Biology Resources and Approaches in Disease Analytics*. London, United Kingdom: IntechOpen doi:10.5772/intechopen.84834

Fernandes, M., Patel, A., and Husi, H. (2018). C/VDdb: A Multi-Omics Expression Profiling Database for a Knowledge-Driven Approach in Cardiovascular Disease (CVD). *PLoS One* 13, e0207371. doi:10.1371/journal.pone.0207371

Fishel, I., Kaufman, A., and Ruppin, E. (2007). Meta-analysis of Gene Expression Data: a Predictor-Based Approach. *Bioinformatics* 23, 1599–1606. doi:10.1093/bioinformatics/btm149

Fisher, R. A. (1992). *Statistical Methods for Research Workers*. New York, NY: Springer, 66–70. doi:10.1007/978-1-4612-4380-9_6

Forero, D. A. (2019). Available Software for Meta-Analyses of Genome-wide Expression Studies. *Cg* 20, 325–331. doi:10.2174/1389202920666190822113912

Furuya, H., Shimizu, Y., and Kawamori, T. (2011). Sphingolipids in Cancer. *Cancer Metastasis Rev.* 30, 567–576. doi:10.1007/s10555-011-9304-1

Fusco, J. P., Pita, G., Pajares, M. J., Andueza, M. P., Patiño-García, A., de-Torres, J. P., et al. (2018). Genomic Characterization of Individuals Presenting Extreme Phenotypes of High and Low Risk to Develop Tobacco-induced Lung Cancer. *Cancer Med.* 7, 3474–3483. doi:10.1002/cam4.1500

Gaj, S., van Erk, A., van Haaften, R. I., and Evelo, C. T. (2007). Linking Microarray Reporters with Protein Functions. *BMC Bioinformatics* 8, 1–17. doi:10.1186/1471-2105-8-360

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: Open Software Development for Computational Biology and Bioinformatics. *Genome Biol.* 5, R80. doi:10.1186/gb-2004-5-10-r80

Goveia, J., Pircher, A., Conradi, L. C., Kalucka, J., Lagani, V., Dewerchin, M., et al. (2016). Meta-analysis of Clinical Metabolic Profiling Studies in Cancer: Challenges and Opportunities. *EMBO Mol. Med.* 8, 1134–1142. doi:10.15252/emmm.201606798

Guan, X., Runger, G., and Liu, L. (2020). Dynamic Incorporation of Prior Knowledge from Multiple Domains in Biomarker Discovery. *BMC Bioinformatics* 21, 77. doi:10.1186/s12859-020-3344-x

Hicks, A. A., Pramstaller, P. P., Johansson, A., Vitart, V., Rudan, I., Ugocsai, P., et al. (2009). Genetic Determinants of Circulating Sphingolipid Concentrations in European Populations. *Plos Genet.* 5, e1000672. doi:10.1371/journal.pgen.1000672

Hong, F., and Breitling, R. (2008). A Comparison of Meta-Analysis Methods for Detecting Differentially Expressed Genes in Microarray Experiments. *Bioinformatics* 24, 374–382. doi:10.1093/BIOINFORMATICS/BTM620

Huan, T., Esko, T., Peters, M. J., Pilling, L. C., Schramm, K., Schurmann, C., et al. (2015). A Meta-Analysis of Gene Expression Signatures of Blood Pressure and Hypertension. *PLOS Genet.* 11, e1005035. doi:10.1371/journal.pgen.1005035

Huang, C., He, C., Ruan, P., and Zhou, R. (2020). TSPYL5 Activates Endoplasmic Reticulum Stress to Inhibit Cell Proliferation, Migration and Invasion in Colorectal Cancer. *Oncol. Rep.* 44, 449–456. doi:10.3892/or.2020.7639

Husi, H. (2004). NMDA Receptors, Neural Pathways, and Protein Interaction Databases. *Int. Rev. Neurobiol.* 61, 49–77. doi:10.1016/S0074-7742(04)61003-8

Jaiswal, A., Gautam, P., Pietilä, E. A., Timonen, S., Nordström, N., Sipari, N., et al. (2020). Multi-modal Meta-Analysis of Cancer Cell Line Omics Profiles Identifies ECHDC1 as a Novel Breast Tumor Suppressor. *bioRxiv* 0131, 929372. doi:10.1101/2020.01.31.929372

Jung, Y., Park, J., Bang, Y.-J., and Kim, T.-Y. (2008). Gene Silencing of TSPYL5 Mediated by Aberrant Promoter Methylation in Gastric Cancers. *Lab. Invest.* 88, 153–160. doi:10.1038/labinvest.3700706

Kale, N. S., Haug, K., Conesa, P., Jayseelan, K., Moreno, P., Rocca-Serra, P., et al. (2016). MetaboLights: An Open-Access Database Repository for Metabolomics Data. *Curr. Protoc. Bioinformatics* 53, 14.13.1. doi:10.1002/0471250953.BI1413S5310.1002/0471250953.bi1413s53

Karczewski, K. J., and Snyder, M. P. (2018). Integrative Omics for Health and Disease. *Nat. Rev. Genet.* 19, 299–310. doi:10.1038/nrg.2018.4

Kim, E. J., Lee, S. Y., Kim, T. R., Choi, S. I., Cho, E. W., Kim, K. C., et al. (2010). TSPYL5 Is Involved in Cell Growth and the Resistance to Radiation in A549 Cells via the Regulation of p21WAF1/Cip1 and PTEN/AKT Pathway. *Biochem. Biophysical Res. Commun.* 392, 448–453. doi:10.1016/j.bbrc.2010.01.045

Kim, S., Jhong, J.-H., Lee, J., and Koo, J.-Y. (2017). Meta-analytic Support Vector Machine for Integrating Multiple Omics Data. *BioData Mining* 10, 1–14. doi:10.1186/s13040-017-0126-8

Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (2018). Meta-analytic Principal Component Analysis in Integrative Omics Application. *Bioinformatics* 34, 1321–1328. doi:10.1093/bioinformatics/btx765

Klimke, W., O'Donovan, C., White, O., Brister, J. R., Clark, K., Fedorov, B., et al. (2011). Solving the Problem: Genome Annotation Standards before the Data Deluge. *Stand. Genomic Sci.* 5, 168–193. doi:10.4056/sigs.2084864

Kröger, W., Mapiye, D., Entfellner, J.-B. D., and Tiffin, N. (2016). A Meta-Analysis of Public Microarray Data Identifies Gene Regulatory Pathways Deregulated in Peripheral Blood Mononuclear Cells from Individuals with Systemic Lupus Erythematosus Compared to Those without. *BMC Med. Genomics* 9, 1–11. doi:10.1186/s12920-016-0227-0

Kutmon, M., van Iersel, M. P., Bohler, A., Kelder, T., Nunes, N., Pico, A. R., et al. (2015). PathVisio 3: An Extendable Pathway Analysis Toolbox. *Plos Comput. Biol.* 11, e1004085. doi:10.1371/journal.pcbi.1004085

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., ur-Rehman, S., et al. (2015). The European Genome-Phenome Archive of Human Data Consented for Biomedical Research. *Nat. Genet.* 47, 692–695. doi:10.1038/ng.3312

Lawson, J., Lee, H., and Lim, Y. (2012). Weighted Geometric Means. *Forum Math.* 24, 1067–1090. doi:10.1515/FORM.2011.096

Li, J., Ma, Z., Shi, M., Malty, R. H., Aoki, H., Minic, Z., et al. (2015). Identification of Human Neuronal Protein Complexes Reveals Biochemical Activities and Convergent Mechanisms of Action in Autism Spectrum Disorders. *Cel Syst.* 1, 361–374. doi:10.1016/j.cels.2015.11.002

Li, X., Qiu, W., Morrow, J., DeMeo, D. L., Weiss, S. T., Fu, Y., et al. (2015). A Comparative Study of Tests for Homogeneity of Variances with Application to DNA Methylation Data. *PLoS One* 10, e0145295. doi:10.1371/JOURNAL.PONE.0145295

Mair, F., Erickson, J. R., Voillet, V., Simoni, Y., Bi, T., Tyznik, A. J., et al. (2020). A Targeted Multi-Omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level. *Cel Rep.* 31, 107499. doi:10.1016/j.celrep.2020.03.063

Maleki, F., Ovens, K., McQuillan, I., and Kusalik, A. J. (2019). Size Matters: How Sample Size Affects the Reproducibility and Specificity of Gene Set Analysis. *Hum. Genomics* 13, 1–12. doi:10.1186/S40246-019-0226-2

McDermott, J. E., Wang, J., Mitchell, H., Webb-Robertson, B.-J., Hafen, R., Ramey, J., et al. (2013). Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert Opin. Med. Diagn.* 7, 37–51. doi:10.1517/17530059.2012.718329

McGarvey, P. B., Nightingale, A., Luo, J., Huang, H., Martin, M. J., Wu, C., et al. (2019). UniProt Genomic Mapping for Deciphering Functional Effects of Missense Variants. *Hum. Mutat.* 40, 694–705. doi:10.1002/humu.23738

Merino, G. A., Fresno, C., Netto, F., Netto, E. D., Pratto, L., and Fernández, E. A. (2016). The Impact of Quality Control in RNA-Seq Experiments. *J. Phys. Conf. Ser.* 705, 012003. Institute of Physics Publishing. doi:10.1088/1742-6596/705/1/012003

Mertins, P., Mani, D. R., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., et al. (2016). Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer. *Nature* 534, 55–62. doi:10.1038/nature18003

Miao, Y., Medeiros, L. J., Xu-Monette, Z. Y., Li, J., and Young, K. H. (2019). Dysregulation of Cell Survival in Diffuse Large B Cell Lymphoma: Mechanisms and Therapeutic Targets. *Front. Oncol.* 9, 107. doi:10.3389/fonc.2019.00107

Miao, Z., and Jiang, X. (2014). Additive and Exclusive Noise Suppression by Iterative Trimmed and Truncated Mean Algorithm. *Signal. Process.* 99, 147–158. doi:10.1016/J.SIGPRO.2013.12.002

Myall, A. C., Perkins, S., Rushton, D., David, J., Spencer, P., Jones, A. R., et al. (2021). An OMICs-Based Meta-Analysis to Support Infection State Stratification. *Bioinformatics* 37, 2347–2355. doi:10.1093/bioinformatics/btab089

Norris, J. L., Farrow, M. A., Gutierrez, D. B., Palmer, L. D., Muszynski, N., Sherrod, S. D., et al. (2017). Integrated, High-Throughput, Multiomics Platform Enables Data-Driven Construction of Cellular Responses and Reveals Global Drug Mechanisms of Action. *J. Proteome Res.* 16, 1364–1375. doi:10.1021/acs.jproteome.6b01004

Pang, Z., Chong, J., Li, S., and Xia, J. (2020). MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. *Metabolites* 10, 186. doi:10.3390/metabo10050186

Parker, B. L., Calkin, A. C., Seldin, M. M., Keating, M. F., Tarling, E. J., Yang, P., et al. (2019). An Integrative Systems Genetic Analysis of Mammalian Lipid Metabolism. *Nature* 567, 187–193. doi:10.1038/s41586-019-0984-y

Perez-Riverol, Y., Bai, M., Da Veiga Leprevost, F., Squizzato, S., Park, Y. M., Haug, K., et al. (2017). Discovering and Linking Public Omics Data Sets Using the Omics Discovery Index. *Nat. Biotechnol.* 35, 406–409. doi:10.1038/nbt.3790

Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (2015). DisGeNET: A Discovery Platform for the Dynamical Exploration of Human Diseases and Their Genes. *Database* 2015, bav028. doi:10.1093/database/bav028

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48, D845–D855. doi:10.1093/nar/gkz1021

Piras, I. S., Manchia, M., Huentelman, M. J., Pinna, F., Zai, C. C., Kennedy, J. L., et al. (2019). Peripheral Biomarkers in Schizophrenia: A Meta-Analysis of Microarray Gene Expression Datasets. *Int. J. Neuropsychopharmacol.* 22, 186–193. doi:10.1093/ijnp/pyy103

Rikke, B. A., Wynes, M. W., Rozeboom, L. M., Barón, A. E., and Hirsch, F. R. (2015). Independent Validation Test of the Vote-Counting Strategy Used to Rank Biomarkers from Published Studies. *Biomarkers Med.* 9, 751–761. doi:10.2217/BMM.15.39

Ringnér, M. (2008). What Is Principal Component Analysis? *Nat. Biotechnol.* 26, 303–304. doi:10.1038/nbt0308-303

Rohart, F., Eslami, A., Matigian, N., Bougeard, S., and Lê Cao, K.-A. (2017). MINT: A Multivariate Integrative Method to Identify Reproducible Molecular Signatures across Independent Experiments and Platforms. *BMC Bioinformatics* 18, 128. doi:10.1186/s12859-017-1553-8

Saha, S., Matthews, D. A., and Bessant, C. (2018). High Throughput Discovery of Protein Variants Using Proteomics Informed by Transcriptomics. *Nucleic Acids Res.* 46, 4893–4902. doi:10.1093/nar/gky295

Saito, M., Gao, J., Basso, K., Kitagawa, Y., Smith, P. M., Bhagat, G., et al. (2007). A Signaling Pathway Mediating Downregulation of BCL6 in Germinal Center B Cells Is Blocked by BCL6 Gene Alterations in B Cell Lymphoma. *Cancer Cell* 12, 280–292. doi:10.1016/j.ccr.2007.08.011

Samaras, P., Schmidt, T., Frejno, M., Gessulat, S., Reinecke, M., Jarzab, A., et al. (2020). ProteomicsDB: a Multi-Omics and Multi-Organism Resource for Life Science Research. *Nucleic Acids Res.* 48, D1153–D1163. doi:10.1093/NAR/GKZ974

Shafi, A., Nguyen, T., Peyvandipour, A., Nguyen, H., and Draghici, S. (2019). A Multi-Cohort and Multi-Omics Meta-Analysis Framework to Identify Network-Based Gene Signatures. *Front. Genet.* 10, 1–16. doi:10.3389/fgene.2019.00159

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Shao, L., Shen, Z., Qian, H., Zhou, S., and Chen, Y. (2017). Knockdown of miR-629 Inhibits Ovarian Cancer Malignant Behaviors by Targeting Testis-specific Y-like Protein 5. *DNA Cel Biol.* 36, 1108–1116. doi:10.1089/dna.2017.3904

Sheppard, E. C., Morrish, R. B., Dillon, M. J., Leyland, R., and Chahwan, R. (2018). Epigenomic Modifications Mediating Antibody Maturation. *Front. Immunol.* 9, 355. doi:10.3389/fimmu.2018.00355

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., et al. (2019). DIABLO: an Integrative Approach for Identifying Key Molecular Drivers from Multi-Omics Assays. *Bioinformatics* 35, 3055–3062. doi:10.1093/bioinformatics/bty1054

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., et al. (2009). BioMart - Biological Queries Made Easy. *BMC Genomics* 10, 1. doi:10.1186/1471-2164-10-22

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 1–25. doi:10.2202/1544-6115.1027

Su, L., Chen, S., Zheng, C., Wei, H., and Song, X. (2019). Meta-Analysis of Gene Expression and Identification of Biological Regulatory Mechanisms in Alzheimer's Disease. *Front. Neurosci.* 13, 633. doi:10.3389/fnins.2019.00633

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* 14, 1177932219899051. doi:10.1177/1177932219899051

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., et al. (2016). Metabolomics Workbench: An International Repository for Metabolomics Data and Metadata, Metabolite Standards, Protocols, Tutorials and Training, and Analysis Tools. *Nucleic Acids Res.* 44, D463–D470. doi:10.1093/NAR/GKV1042

Todorov, H., Fournier, D., and Gerber, S. (2018). Principal Components Analysis: Theory and Application to Gene Expression Data Analysis. *Genomics Comput. Biol.* 4, 100041. doi:10.18547/gcb.2018.vol4.iss2.e100041

Toro-Domínguez, D., Villatoro-García, J. A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M. E., and Carmona-Sáez, P. (2020). A Survey of Gene Expression Meta-Analysis: Methods and Applications. *Brief. Bioinform* 22, 1694–1705. doi:10.1093/bib/bbaa019

Tsuyama, N., Sakata, S., Baba, S., Mishima, Y., Nishimura, N., Ueda, K., et al. (2017). BCL2 Expression in DLBCL: Reappraisal of Immunohistochemistry with New Criteria for Therapeutic Biomarker Evaluation. *Blood* 130, 489–500. doi:10.1182/blood-2016-12-759621

Vachani, A., Nebozhyn, M., Singhal, S., Alila, L., Wakeam, E., Muschel, R., et al. (2007). A 10-gene Classifier for Distinguishing Head and Neck Squamous Cell Carcinoma and Lung Squamous Cell Carcinoma. *Clin. Cancer Res.* 13, 2905–2915. doi:10.1158/1078-0432.CCR-06-1670

van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., et al. (2010). The BridgeDb Framework: Standardized Access to Gene, Protein and Metabolite Identifier Mapping Services. *BMC Bioinformatics* 11, 5. doi:10.1186/1471-2105-11-5

Vennou, K. E., Piovani, D., Kontou, P. I., Bonovas, S., and Bagos, P. G. (2020). Methods for Multiple Outcome Meta-Analysis of Gene-Expression Data. *MethodsX* 7, 100834. doi:10.1016/j.mex.2020.100834

Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A., Fabregat, A., Foster, J. M., et al. (2013). The Proteomics Identifications (PRIDE) Database and Associated Tools: Status in 2013. *Nucleic Acids Res.* 41, D1063–D1069. doi:10.1093/NAR/GKS1262

Waldron, L., and Riester, M. (2016). Meta-analysis in Gene Expression Studies. *Methods Mol. Biol.* 1418, 161–176. doi:10.1007/978-1-4939-3578-9_8

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., et al. (2016). Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* 34, 828–837. doi:10.1038/NBT.3597

Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W., and Bandeira, N. (2018). Assembling the Community-Scale Discoverable Human Proteome. *Cel Syst.* 7, 412–421. doi:10.1016/J.CELS.2018.08.004

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., et al. (2012). PaxDb, a Database of Protein Abundance Averages across All Three Domains of Life. *Mol. Cell Proteomics* 11, 492–500. doi:10.1074/MCP.O111.014704

Wang, Q., Li, W.-X., Dai, S.-X., Guo, Y.-C., Han, F.-F., Zheng, J.-J., et al. (2017). Meta-Analysis of Parkinson's Disease and Alzheimer's Disease Revealed Commonly Impaired Pathways and Dysregulation of NRF2-dependent Genes. *Jad* 56, 1525–1539. doi:10.3233/JAD-161032

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function. *Nucleic Acids Res.* 38, W214–W220. Available at: https://academic.oup.com/nar/article/38/suppl_2/W214/1126704 (Accessed September 24, 2020). doi:10.1093/nar/gkq537

Watanabe, Y., Yoshizawa, A. C., Ishihama, Y., and Okuda, S. (2021). The jPOST Repository as a Public Data Repository for. *Methods Mol. Biol.* 2259, 309–322. doi:10.1007/978-1-0716-1178-4_20

Webster, J. D., and Vucic, D. (2020). The Balance of TNF Mediated Pathways Regulates Inflammatory Cell Death Signaling in Healthy and Diseased Tissues. *Front. Cel Dev. Biol.* 8, 365. doi:10.3389/fcell.2020.00365

Winter, C., Kosch, R., Ludlow, M., Osterhaus, A. D. M. E., and Jung, K. (2019). Network Meta-Analysis Correlates with Analysis of Merged Independent Transcriptome Expression Data. *BMC Bioinformatics* 20, 1–10. doi:10.1186/s12859-019-2705-9

Xia, J., Fjell, C. D., Mayer, M. L., Pena, O. M., Wishart, D. S., and Hancock, R. E. W. (2013). INMEX-a Web-Based Tool for Integrative Meta-Analysis of Expression Data. *Nucleic Acids Res.* 41, W63–W70. doi:10.1093/nar/gkt338

Xicota, L., Ichou, F., Lejeune, F.-X., Colsch, B., Tenenhaus, A., Leroy, I., et al. (2019). Multi-omics Signature of Brain Amyloid Deposition in Asymptomatic Individuals At-Risk for Alzheimer's Disease: The INSIGHT-preAD Study. *EBioMedicine* 47, 518–528. doi:10.1016/j.ebiom.2019.08.051

Xu, W., Dullaers, M., Oh, S., and Banchereau, J. (2009). Distinct Roles of Dendritic Cells and Macrophages in B Cell Class Switching (39.12). *J. Immunol.* 182, 39.

Yan, J., Risacher, S. L., Shen, L., and Saykin, A. J. (2017). Network Approaches to Systems Biology Analysis of Complex Disease: Integrative Methods for Multi-Omics Data. *Brief. Bioinform.* 19, 1370–1381. doi:10.1093/bib/bbx066

Yang, X. (2020). Multitissue Multiomics Systems Biology to Dissect Complex Diseases. *Trends Mol. Med.* 26, 718–728. doi:10.1016/j.molmed.2020.04.006

Zaykin, D. V., Zhivotovsky, L. A., Czika, W., Shao, S., and Wolfinger, R. D. (2007). Combining P -values in Large-Scale Genomics Experiments. *Pharmaceut. Statist.* 6, 217–226. doi:10.1002/PST.304

Zhang, B., Calado, D. P., Wang, Z., Fröhler, S., Köchert, K., Qian, Y., et al. (2015). An Oncogenic Role for Alternative NF-Kb Signaling in DLBCL Revealed upon Deregulated BCL6 Expression. *Cel Rep.* 11, 715–726. doi:10.1016/j.celrep.2015.03.059

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., et al. (2013). Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell* 153, 707–720. doi:10.1016/j.cell.2013.03.030

Zhou, G., Li, S., and Xia, J. (2020). Network-Based Approaches for Multi-Omics Integration. *Methods Mol. Biol.* 2104, 469–487. doi:10.1007/978-1-0716-0239-3_23

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., and Xia, J. (2019). NetworkAnalyst 3.0: A Visual Analytics Platform for Comprehensive Gene Expression Profiling and Meta-Analysis. *Nucleic Acids Res.* 47, W234–W241. doi:10.1093/nar/gkz240

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 March 2022
doi: 10.3389/fgene.2022.814412

Integrative Pathway Analysis of SNP and Metabolite Data Using a Hierarchical Structural Component Model

Taeyeong Jung[1], Youngae Jung[2], Min Kyong Moon[3], Oran Kwon[4], Geum-Sook Hwang[2] and Taesung Park[1,5]**

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea, [2]Korea Integrated Metabolomics Research Group, Western Seoul Center, Korea Basic Science Institute, Seoul, South Korea, [3]Department of Internal Medicine, Seoul National University Boramae Medical Center, Seoul, South Korea, [4]Department of Nutritional Science and Food Management, Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul, South Korea, [5]Department of Statistics, Seoul National University, Seoul, South Korea

Integrative multi-omics analysis has become a useful tool to understand molecular mechanisms and drug discovery for treatment. Especially, the couplings of genetics to metabolomics have been performed to identify the associations between SNP and metabolite. However, while the importance of integrative pathway analysis is increasing, there are few approaches to utilize pathway information to analyze phenotypes using SNP and metabolite. We propose an integrative pathway analysis of SNP and metabolite data using a hierarchical structural component model considering the structural relationships of SNPs, metabolites, pathways, and phenotypes. The proposed method utilizes genome-wide association studies on metabolites and constructs the genetic risk scores for metabolites referred to as genetic metabolomic scores. It is based on the hierarchical model using the genetic metabolomic scores and pathways. Furthermore, this method adopts a ridge penalty to consider the correlations between genetic metabolomic scores and between pathways. We apply our method to the SNP and metabolite data from the Korean population to identify pathways associated with type 2 diabetes (T2D). Through this application, we identified well-known pathways associated with T2D, demonstrating that this method adds biological insights into disease-related pathways using genetic predispositions of metabolites.

Keywords: pathway analysis, multi-omics integration, mGWAS, metabolite, SNP

OPEN ACCESS

Edited by:
Miguel E. Renteria, QIMR Berghofer Medical Research Institute, Australia

Reviewed by:
Jung Hun Oh, Memorial Sloan Kettering Cancer Center, United States
Jianguo Xia, McGill University, Canada

****Correspondence:***
Geum-Sook Hwang
gshwang@kbsi.re.kr
Taesung Park
tspark@stats.snu.ac.kr

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *13 November 2021*
Accepted: *13 January 2022*
Published: *24 March 2022*

Citation:
Jung T, Jung Y, Moon MK, Kwon O, Hwang G-S and Park T (2022) Integrative Pathway Analysis of SNP and Metabolite Data Using a Hierarchical Structural Component Model. Front. Genet. 13:814412. doi: 10.3389/fgene.2022.814412

1 INTRODUCTION

The advances in biological techniques have led to the generation of multiple omics (multi-omics) data, which contribute to a better understanding of biological mechanisms and diseases. For instance, the next-generation sequencing (NGS) technology for genome-wide data and mass spectrometry for quantitative metabolic data allow us to generate multi-omics data from the same samples at a low cost (Metzker, 2010; Suhre and Gieger, 2012). These technical improvements have enabled multi-omics data analysis to become a useful tool in biomedical research.

Genome-wide association studies (GWAS) have been conducted worldwide to identify single nucleotide polymorphisms (SNPs) associated with various diseases or phenotypes.

An intermediate variable, linking genetic variants and phenotype, is suggested to consider the effects of genes and environmental factors in overcoming the limitation of GWAS (Kronenberg, 2012). One of the potential intermediate variables is serum metabolite concentration, providing a direct readout of biological processes, to connect genetic factors and diseases (Illig et al., 2010; Kronenberg, 2012). Recently, metabolite genome-wide association studies (mGWAS) and metabolic quantitative trait loci (mQTL) analyses have been conducted by utilizing SNP and metabolite data together (Zhang et al., 2017; Park et al., 2019; Ouyang et al., 2021). In addition, to explore the association between SNPs and metabolites, disease-related metabolomic markers using SNPs were investigated through Mendelian randomization (Moayyeri et al., 2018). Even though many studies attempted to analyze SNP and metabolite data together, most studies have mainly focused on either analyzing statistical associations between SNPs and metabolites or discovering metabolomic markers of phenotypes using SNPs.

Since pathway analysis can give a more intuitive interpretation of the biological system, several methods have been proposed for pathway analysis that focuses on identifying significant pathways related to certain traits of interest (García-Campos et al., 2015; Kao et al., 2017). Specifically, pathway analysis using multi-omics data has now become popularly used in recent bioinformatics research. While the importance of integrative pathway analysis is increasing, there have been few studies about integrating SNPs and metabolite data (Kao et al., 2017). In this study, we focus on integrative pathway analysis of SNPs and metabolite data.

Here, we propose an integrative pathway analysis of SNP and metabolite data using a hierarchical structural component model. This method calculates genetic risk scores of metabolites and investigates pathways associated with phenotypes through the genetic risk scores. This approach is based on our earlier work Pathway-based approach using HierArchical components of collapsed RAre variants Of High-throughput sequencing data (PHARAOH) (Lee et al., 2016). PHARAOH uses rare variants to construct collapsed genes and performs pathway analysis using these gene-summaries. PHARAOH simultaneously analyzes the entire collapsed genes and the entire pathways in a hierarchical model (Lee et al., 2016). We utilize this main framework of PHARAOH and mGWAS for the integration of SNP and metabolite data and refer to this method as a Hierarchical Structural Component Model of SNP and Metabolite data for pathway analysis (HisCoM-SM).

The genetic metabolomic score (GMS) is calculated by summing the effects of the corresponding SNPs on each metabolite and then is used for pathway analysis in PHARAOH. HisCoM-SM adopts the ridge penalties to both GMSs and pathways to identify pathways while controlling for potential correlations between GMSs and between pathways.

Here, we apply HisCoM-SM to SNP and metabolite data from Korean Association REsource (KARE) cohort to identify pathways associated with T2D. Note that T2D is a metabolic disorder that is affected by genetic factors and environmental exposure simultaneously (Murea et al., 2012). Through this application to the KARE dataset, we demonstrate that HisCoM-SM can identify previously reported pathways, including insulin secretion and insulin resistance, associated with T2D, using genetic predispositions of metabolites (Weyer et al., 2001; Dayeh et al., 2014; Kahn et al., 2014).

The HisCoM-SM is available at https://statgen.snu.ac.kr/software/HisCoM_SM.

TABLE 1 | Number of metabolites in each category.

Category	Number of metabolites
Alkaloids and derivatives	1
Benzenoids	2
Lipids and lipid-like molecules	1
Nucleosides, nucleotides, and analogues	4
Organic acids and derivatives	33
Organic nitrogen compounds	4
Organic oxygen compounds	1
Organoheterocyclic compounds	7

2 MATERIALS AND METHODS

2.1 SNP Data

The SNP data was generated by the Affymetrix Genome-Wide Human SNP array 5.0. from the Korea Association REsource (KARE) project. KARE is based on Ansan and Ansung Korean population cohort among 10,038 participants which was initiated in 2001 (Cho et al., 2009). This chip originally consisted of 8,840 individuals and 352,228 SNPs. We applied quality control to our SNP data to reduce the biases and used common variants for our analysis (Turner et al., 2011). For quality control of SNP data, the genotypes with over 0.1 missing rates and Hardy-Weinberg equilibrium p-values $< 10^{-6}$ were excluded. To use only common variants, the genotypes with minor allele frequency (MAF) $\leq$ 0.05 were excluded. Then, we retained the individuals who have metabolite data and whose calling rate >0.9. After quality control of SNPs from the KARE dataset using PLINK 1.90, a total of 312,116 SNPs were analyzed in this work (Chang et al., 2015).

2.2 Metabolite Data

The serum metabolites in the 691 participants were quantitatively analyzed by a targeted metabolomics approach using liquid chromatography-mass spectrometry (LC-MS). 64 metabolites were measured in this work. The metabolites of each subject were measured at the fifth follow-up in the KARE dataset. Among 64 metabolites, 53 were mapped to 101 pathways. The 53 metabolites were classified into eight categories. **Table 1** shows the number of metabolites in each category. The list of metabolites and the eight categories of metabolites are shown in **Supplementary Table S1**. 627 samples were available with both SNPs and metabolite data. Among these samples, 309 samples are controls (normal) and 318 samples are cases (pre-T2D and T2D). For metabolite data, systematical error removal using random forest (SERRF) was used for batch effect correction to remove variation due to instrument and injection time (Fan et al., 2019).

TABLE 2 | The characteristics of the subjects in each case (pre T2D + T2D) and control (Normal) group.

	Case	Control	p-value
Male	157 (49.37%)	157 (50.81%)	0.7794
Age (years)	58.26	57.32	0.0653
BMI	25.22	24.60	0.0059
Number of subjects	318	309	—

2.3 Diagnosis of Type 2 Diabetes

The criteria for diagnosis of T2D are 1) fasting blood glucose (FBS) ≥ 126 mg/dl, 2) 2-h postprandial glucose (2 PP) ≥ 200 mg/dl, 3) HbA1c ≥ 6.5 (%), and 4) treatment of drug. Pre-diabetic (preT2D) individuals are diagnosed by the criteria—1) 100 mg/dl ≤ FBS < 126 mg/dl, 2) 140 mg/dl ≤ 2 PP < 200 mg/dl, 3) 5.6% < HbA1c < 6.5%, and 4) no treatment of drug. The criteria for normal individual are 1) FBS <100, 2) 2 PP < 140, 3) HbA1c ≤ 5.6%, and 4) no treatment of drug. Here, we regarded T2D and preT2D individuals as cases, and normal individuals as controls. The baseline characteristics of those samples are shown in **Table 2**.

2.4 HisCoM-SM

The framework of HisCoM-SM consists of two steps. The step 1 is to calculate genetic risk scores of metabolite data referred to genetic metabolomic scores (GMSs). Genetic effects of metabolites are estimated and then used to calculate the GMSs. The step 2 is to perform pathway analysis using the calculated GMSs by step 1. To perform pathway analysis, a hierarchical structural component model (HisCoM) is used. HisCoM consists of three layers which are input layer, latent layer, and outcome layer. In our work, GMSs are used as input variables, pathways are used as latent variables, and binary phenotype is used as outcome variable. The two steps of the process are described in more detail below.

2.4.1 Calculation of GMSs

To perform pathway analysis using SNP and metabolite data, we first construct the GMSs. Here, we used two methods for calculating GMSs. The first score was derived from the single-SNP association test for metabolites. To do that, we applied linear regression for each metabolite adjusted for age and sex and calculated the GMSs by clumping and thresholding to remove redundant correlated effects due to linkage disequilibrium (LD) using PLINK (Chang et al., 2015). Clumping is the process of selecting the most significant SNP iteratively, computing correlation between this SNP and nearby SNPs within a genetic distance of 250 k, and removing all the nearby SNPs with highly correlation ($r^2 > 0.2$) (Privé et al., 2019). Thresholding is the process of filtering out variants with p-values greater than a given threshold level (Privé et al., 2019). Then, the GMSs are calculated from the effects of remaining SNPs after clumping and thresholding using PLINK (Chang et al., 2015).

The second score is based on the genetic best linear unbiased prediction (GBLUP) method from Genome-wide Complex Trait Analysis (GCTA) software (Yang et al., 2011). All SNPs are treated as random effects in a mixed linear model adjusted for fixed effects of sex and age (Yang et al., 2011). In GBLUP, the effects of all SNPs are estimated by the genetic relationship matrix (GRM) representing the relatedness of individuals' SNPs (Yang et al., 2011). The GRM is used to estimate the effects of all SNPs and only 20% of SNPs with a high absolute value of the effect are used to construct the GMSs. Then, the remaining SNP effects are used to construct GMSs using PLINK (Chang et al., 2015).

2.4.2 Pathway Analysis Using GMSs in a Hierarchical Component Model

After constructing the GMSs, pathway analysis is performed. **Figure 1** shows the diagram of HisCoM-SM. For each metabolite, the correlated SNPs are selected by a single SNP association test and GBLUP. Then, GMSs are derived as a linear combination of these SNPs. Thus, each metabolite is linearly correlated with the selected multiple SNPs. Similarly, each pathway is also linearly correlated with multiple metabolites. First, an individual pathway is mapped to the metabolites using the KEGG pathway database. Next, the latent variables representing pathways are derived as a linear combination of these metabolites. Then, the binary outcome is used to estimate the effect of the relationship between pathways and the phenotype. Let y_j be the binary outcome of the j^{th} individual, K be the number of pathways, T_k be the number of GMSs in the k^{th} pathway. The x_{jkt} denotes GMS which is a continuous value, the w_{kt} represents weight for x_{jkt}, and β_k denotes the coefficient for pathway. Then, the proposed HisCoM model is defined as follows:

$$logit(\pi_j) = \beta_0 + \sum_{k=1}^{K}\left[\sum_{t=1}^{T_k} w_{kt}x_{jkt}\right]\beta_k \quad (1)$$

To estimate the parameters of the model, we maximize a penalized log-likelihood function (**Eq. 2**) and use alternating least squares (ALS) for minimizing the objective function (Lee et al., 2016). Let $p(y_j; \gamma_j, \delta)$ be the probability distribution function for the phenotype y_j, λ_M and λ_P denote ridge parameters added for potential multicollinearities between GMSs and between pathways, respectively. After determining the ridge parameters λ_M and λ_P by five-fold cross-validation, the coefficients w_{kt} and β_k are estimated by ALS algorithm. In ALS algorithm, β_k are updated in a least square manner with w_{kt} fixed. Likewise, w_{kt} are

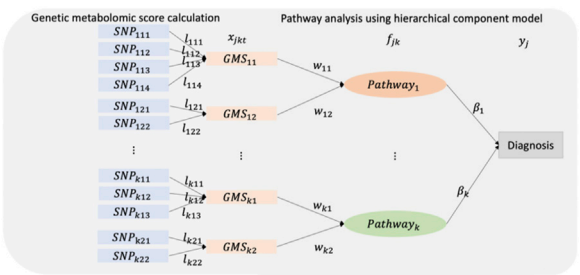

FIGURE 1 | A schematic diagram of the HisCoM-SM.

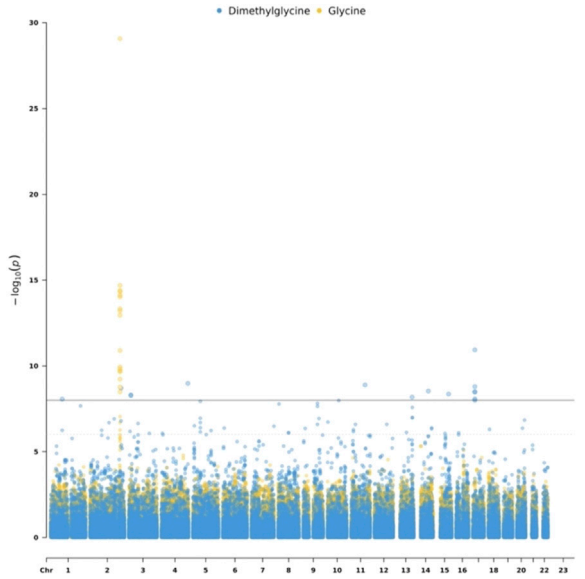

FIGURE 2 | Manhattan plot for Dimethylglycine and Glycine

updated with β_k fixed. This ALS algorithm is iterated until convergence.

$$\phi = \sum_{j=1}^{N} log\ p\left(y_j; \gamma_j, \delta\right) - 1/2\lambda_M \sum_{k=1}^{K} \sum_{t=1}^{T_k} w_{kt}^2 - 1/2\lambda_P \sum_{k=0}^{K} \beta_k^2 \quad (2)$$

After estimation, the phenotype is resampled 100,000 times through permutation to generate the null distribution of coefficients of pathways to calculate empirical p-values. To correct the multiple comparisons problem, the false discovery rate (FDR) is applied (Benjamini and Hochberg, 1995). Here, we use the WISARD (workbench for integrated superfast association studies for related datasets) to perform integrative pathway analysis using GMSs (Lee et al., 2018).

3 RESULTS

3.1 Metabolite Genome-wide Association Study in KARE Dataset

To detect genetic variants associated with metabolites, we performed the mGWAS using linear regression, adjusting for sex and age. Out of 53 metabolites, only two metabolites have significantly ($p < 1e\text{-}8$) associated SNPs. Specifically, we identified 17 SNPs associated with Glycine which is related to insulin sensitivity and secretion (Floegel et al., 2013). These SNPs are located on chromosome 2. We also identified 15 SNPs associated with Dimethylglycine. The list of the identified mQTL is shown in **Supplementary Table S2**. **Figure 2** is a Manhattan plot for SNPs associated with Glycine and Dimethylglycine.

3.2 Pathway Analysis of T2D

HisCoM-SM was applied to SNP and metabolite data of T2D/preT2D patients and normal samples from the KARE dataset which is a large Korean population-based cohort. We first mapped the KEGG pathway database with metabolite data. Among 64 metabolites, 53 metabolites were mapped to 101 pathways. Then, the GMSs were used as components of pathways, which are latent variables in the model. Note that we used the two methods to construct GMSs: 1) Single-SNP association-based GMSs and 2) GBLUP-based GMSs. Those two methods are discussed in detail in the Methods section. The lists of identified pathways with HisCoM-SM based on single-SNP association denoted by HisCoM-SM (single) and HisCoM-SM based on GBLUP denoted by HisCoM-SM(GBLUP) are shown in **Supplementary Tables S3, S4**, respectively.

To detect the significant pathways associated with T2D in HisCoM-SM, we selected the pathways with high absolute coefficient values and low q-values. The metabolic pathway (map 01100) had the highest absolute effect value and the lowest q-value in both HisCoM-SM (single) and HisCoM-SM(GBLUP). Among the 49 metabolites contained in this pathway, Arginine, Tryptophan, Lactate, Trimethylamine N-oxide (TMAO), *Trans*-4-Hydroxy-L-proline, and Hippurate were significant in both HisCoM-SM methods. Arginine facilitates the action of glucose to stimulate insulin release (Gerich et al., 1974). In addition to Arginine, the other five metabolites have also been reported as risk factors for the incidence of T2D or the prevalence of T2D (Van Doorn et al., 2007; Crawford et al., 2010; Chen et al., 2016; Shan et al., 2017; Tang et al., 2017). In addition, both HisCoM-SM methods identified the same pathway with the second-highest absolute coefficient value and the lowest q-value. This pathway is the biosynthesis of amino acids (map 01230) and has also been reported to be associated with T2D in previous studies (Aichler et al., 2017; Lu et al., 2019). These results demonstrate that HisCoM-SM(single) and HisCoM-SM(GBLUP) can yield consistent results and identify pathways associated with T2D.

3.3 Comparison of HisCoM-SM to Conventional HisCoM Using Metabolite Data

For comparison purposes, we applied the conventional HisCoM to KARE metabolite data to identify the T2D related pathways (**Supplementary Table S5**). **Table 3** summarizes the commonly significant (FDR q-value < 0.05) pathways by HisCoM-SM(single), HisCoM-SM(GBLUP), and conventional HisCoM using only metabolite data. These commonly significant pathways are categorized by the KEGG pathway category and subcategory. Metabolism is the category that has the greatest number of significant pathways. Among the 64 significant pathways, 31 pathways are included in the metabolism category. **Figure 3** is a scatter plot for the FDR q-values and the correlation coefficients for each pair of methods. Here, the q-values of HisCoM-SM and HisCoM showed quite consistent patterns yielding high correlation coefficients. **Figure 4** is a Venn diagram to show the numbers of significant pathways (FDR q-value <

TABLE 3 | Identified common pathways in HisCoM-SM and conventional HisCoM (q-value < 0.05). The pathways are categorized by KEGG pathway categories and KEGG pathway subcategories. The values in parenthesis are the number of pathways included in the KEGG pathway.

KEGG pathway category	KEGG pathway subcategory	Pathway
Cellular Processes (3)	Cell growth and death	Ferroptosis
	Cell motility	Regulation of actin cytoskeleton
	Cellular community - eukaryotes	Gap junction
Environmental Information Processing (4)	Membrane transport	ABC transporters
	Signal transduction	mTOR signaling pathway/Sphingolipid signaling pathway
	Signaling molecules and interaction	Neuroactive ligand-receptor interaction
Genetic Information Processing (2)	Folding, sorting, and degradation	Sulfur relay system
	Translation	Aminoacy-tRNA biosynthesis
Human Diseases (9)	Drug resistance: antineoplastic	Antifolate resistance
	Endocrine and metabolic disease	Insulin resistance
	Neurodegenerative disease	Amyotrophic lateral sclerosis/Parkinson disease
	Substance dependence	Alcoholism/Amphetamine addiction/cocaine addiction/Morphine addiction/Nicotine addiction
Metabolism (31)	Amino acid metabolism	Alanine, aspartate and glutamate metabolism/Arginine and proline metabolism/Arginine biosynthesis/Cysteine and methionine metabolism/Glycine, serine and threonine metabolism/Histidine metabolism/Phenylalanine metabolism/Phenylalanine, tyrosine and tryptophan biosynthesis/Tyrosine metabolism/Valine, leucine and isoleucine biosynthesis/Valine, leucine, and isoleucine degradation
	Biosynthesis of other secondary metabolites	Caffeine metabolism/Neomycin, kanamycin, and gentamicin biosynthesis
	Carbohydrate metabolism	Butanoate metabolism/Glyoxylate and dicarboxylate metabolism/Pyruvate metabolism
	Energy metabolism	Nitrogen metabolism
	Global overview maps	2-Oxocarboxylic acid metabolism/Biosynthesis of amino acids/Carbon metabolism/Metabolic pathways
	Metabolism of cofactors and vitamins	Nicotinate and nicotinamide metabolism/Pantothenate and CoA biosynthesis/Porphyrin and chlorophyll metabolism/Thiamine metabolism
	Metabolism of other amino acids	beta-Alanine metabolism/D-Arginine and D-ornithine metabolism/D-glutamine and D-glutamate metabolism/Glutathione metabolism/Taurine and hypotaurine metabolism
	Nucleotide metabolism	Purine metabolism
Organismal Systems (15)	Digestive system	Bile secretion/Mineral absorption/Pancreatic secretion/Protein digestion and absorption
	Endocrine system	Estrogen signaling pathway/Insulin secretion/Prolactin signaling pathway
	Excretory system	Proximal tubule bicarbonate reclamation
	Nervous system	Dopaminergic synapse/GABAergic synapse/Glutamatergic synapse/Long-term depression/Retrograde endocannabinoid signaling/Synaptic vesicle cycle
	Sensory system	Taste transduction

0.05) shared by different methods. Note that 64 out of 74 significant pathways were commonly identified by all three methods, indicating that HisCoM based methods yielded quite consistent results. Also, all pathways identified by HisCoM-SM (single) were identified by HisCoM-SM(GBLUP).

In **Figure 4**, conventional HisCoM identified two pathways that HisCoM-SM could not detect, one of which (selenocompound metabolism; map 00450) was previously reported to be associated with T2D (Shin et al., 2020). On the other hand, HisCoM-SM(GBLUP) identified three pathways, which conventional HisCoM could not find. Two out of these pathways were reported as associated with T2D. These two significant pathways are biotin metabolism (map 00780) and vascular smooth muscle contraction (map 04270) (Xie et al., 2006; Hashimoto et al., 2020). For biotin metabolism, several studies have shown that plasma triacylglycerol, low-density lipoprotein cholesterol (LDL), and fasting glucose are reduced in patients with T2D who take biotin supplementation (Maebashi et al., 1993; Revilla-Monsalve et al., 2006). Furthermore, biotin intake has been reported to be effective in improving glycaemic control through diabetic animal models (Reddi et al., 1988; Zhang et al., 1997).

4 DISCUSSION

Several studies have suggested that pathway analysis using multi-omics data allows more insights into biological systems. Pathway analysis using more than one omics data is becoming increasingly common. However, few studies can identify disease-related pathways considering SNPs and metabolites together.

We proposed a novel pathway analysis integrating SNP and metabolite data. Our method introduced novel genetic metabolomic scores (GMSs) for pathway analysis. We used a single-SNP association and a GBLUP approach to construct GMSs. The calculated GMSs were used as components of pathways in a hierarchical model. The coefficients can be estimated by analyzing GMSs and pathways simultaneously,

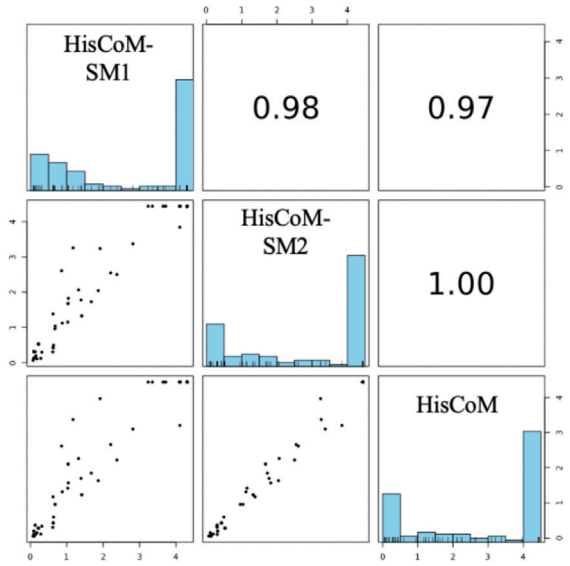

FIGURE 3 | Pairwise scatter plot of 101 pathways' FDR q-values calculated by HisCoM-SM methods and HisCoM using metabolite. Note that His-CoM-SM1 represents HisCoM-SM(single) and HisCoM-SM2 represents HisCoM-SM(GBLUP).

considering the correlations between these scores and between pathways, respectively.

We applied HisCoM-SM to the KARE cohort dataset. Our HisCoM-SM successfully identified pathways that were reported to be related to T2D. In our result, the pathways identified by HisCoM-SM and conventional HisCoM were almost overlapped, indicating that HisCoM-SM and HisCoM yielded quite consistent results, and the GMSs can be utilized for pathway analysis. Moreover, HisCoM-SM could identify the T2D-related pathways that the conventional HisCoM using only metabolite data could not detect. Since 53 targeted metabolomics in our analysis may cover only a small portion of the metabolome, modeling the effects of SNPs on these metabolites resulted in similar results from the conventional HisCoM method using only metabolites. We are planning to modify HisCoM-SM so that it allows for each pathway to have inputs from both genes and metabolites simultaneously. In other words, SNPs can directly contribute to pathways (not through metabolites), which also makes a more biological sense. The new model with a rewired structure is expected to improve the performance. We will leave it as a near-future study.

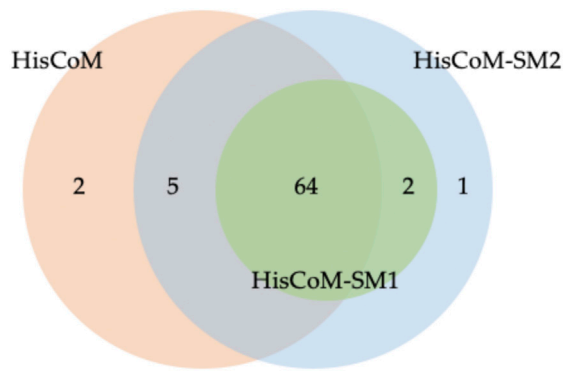

FIGURE 4 | Venn Diagram for numbers of significant pathways detected for each method. Note that HisCoM-SM1 represents HisCoM-SM(single) and HisCoM-SM2 represents HisCoM-SM(GBLUP).

Here, we applied the clumping and thresholding process to generate genetic metabolomic scores using p-values from linear regression models. Instead of linear regression models, other approaches such as Kernel regression can be applied to detect non-linear relationships between SNPs and metabolites. Our HisCoM-SM can also use other GMSs such as the ones derived from the LD pred method (Vilhjálmsson et al., 2015). Also, once the effect of SNPs on each metabolite is obtained, it can be used to calculate the GMSs for other datasets only with SNPs. The GMSs can be calculated using the effects of SNPs. Regarding the estimation of effects of pathways and genetic metabolomic scores, we can use different types of penalty functions. For example, LASSO or Elastic Net can be easily incorporated into our model instead of the Ridge penalty. Furthermore, we can construct a predictive model using HisCoM-SM approach for diagnosis. Specifically, we will evaluate the prediction performance of HisCoM-SM and compare it with those of other models such as original HisCoM using only SNPs and metabolites in a near future.

We believe that our method may add practical biological insights into the disease-related pathways by genetic predispositions of metabolites and contribute to the understanding of molecular mechanisms and treatment for the disease.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://koreabiobank.re.kr.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Seoul National University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

TJ and TP contributed to conception and design of the study. YJ and G-SH generated the metabolite data. OK acquisition the funding. TJ and TP designed the statistical model and TJ performed the analysis. TJ wrote the first draft of the manuscript and TP edited the draft. YJ wrote sections of the

manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was funded by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation, the Korea Health Technology R8D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare (HI16C2037), and the Korea Basic Science Institute (270000).

ACKNOWLEDGMENTS

The author would like to thank Apio Catherine for the comments to edit this manuscript for English.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.814412/full#supplementary-material

REFERENCES

Aichler, M., Borgmann, D., Krumsiek, J., Buck, A., Macdonald, P. E., Fox, J. E. M., et al. (2017). N-acyl Taurines and Acylcarnitines Cause an Imbalance in Insulin Synthesis and Secretion Provoking β Cell Dysfunction in Type 2 Diabetes. *Cel. Metab.* 25, 1334–1347. e1334. doi:10.1016/j.cmet.2017.04.012

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of Larger and Richer Datasets. *Gigascience* 4, 7. doi:10.1186/s13742-015-0047-8

Chen, T., Zheng, X., Ma, X., Bao, Y., Ni, Y., Hu, C., et al. (2016). Tryptophan Predicts the Risk for Future Type 2 Diabetes. *PloS one* 11, e0162192. doi:10.1371/journal.pone.0162192

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H.-J., et al. (2009). A Large-Scale Genome-wide Association Study of Asian Populations Uncovers Genetic Factors Influencing Eight Quantitative Traits. *Nat. Genet.* 41, 527–534. doi:10.1038/ng.357

Crawford, S. O., Hoogeveen, R. C., Brancati, F. L., Astor, B. C., Ballantyne, C. M., Schmidt, M. I., et al. (2010). Association of Blood Lactate with Type 2 Diabetes: the Atherosclerosis Risk in Communities Carotid MRI Study. *Int. J. Epidemiol.* 39, 1647–1655. doi:10.1093/ije/dyq126

Dayeh, T., Volkov, P., Salö, S., Hall, E., Nilsson, E., Olsson, A. H., et al. (2014). Genome-Wide DNA Methylation Analysis of Human Pancreatic Islets from Type 2 Diabetic and Non-diabetic Donors Identifies Candidate Genes that Influence Insulin Secretion. *Plos Genet.* 10, e1004160. doi:10.1371/journal.pgen.1004160

Fan, S., Kind, T., Cajka, T., Hazen, S. L., Tang, W. H. W., Kaddurah-Daouk, R., et al. (2019). Systematic Error Removal Using Random forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal. Chem.* 91, 3590–3596. doi:10.1021/acs.analchem.8b05592

Floegel, A., Stefan, N., Yu, Z., Mühlenbruch, K., Drogan, D., Joost, H.-G., et al. (2013). Identification of Serum Metabolites Associated with Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. *Diabetes* 62, 639–648. doi:10.2337/db12-0495

García-Campos, M. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2015). Pathway Analysis: State of the Art. *Front. Physiol.* 6, 383. doi:10.3389/fphys.2015.00383

Gerich, J. E., Charles, M. A., and Grodsky, G. M. (1974). Characterization of the Effects of Arginine and Glucose on Glucagon and Insulin Release from the Perfused Rat Pancreas. *J. Clin. Invest.* 54, 833–841. doi:10.1172/jci107823

Hashimoto, Y., Hamaguchi, M., Kaji, A., Sakai, R., Osaka, T., Inoue, R., et al. (2020). Intake of Sucrose Affects Gut Dysbiosis in Patients with Type 2 Diabetes. *J. Diabetes Investig.* 11, 1623–1634. doi:10.1111/jdi.13293

Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., et al. (2010). A Genome-wide Perspective of Genetic Variation in Human Metabolism. *Nat. Genet.* 42, 137–141. doi:10.1038/ng.507

Kahn, S. E., Cooper, M. E., and Del Prato, S. (2014). Pathophysiology and Treatment of Type 2 Diabetes: Perspectives on the Past, Present, and Future. *Lancet* 383, 1068–1083. doi:10.1016/s0140-6736(13)62154-6

Kao, P. Y. P., Leung, K. H., Chan, L. W. C., Yip, S. P., and Yap, M. K. H. (2017). Pathway Analysis of Complex Diseases for GWAS, Extending to Consider Rare Variants, Multi-Omics and Interactions. *Biochim. Biophys. Acta (Bba) - Gen. Subj.* 1861, 335–353. doi:10.1016/j.bbagen.2016.11.030

Kronenberg, F. (2012). "Metabolic Traits as Intermediate Phenotypes," in *Genetics Meets Metabolomics* (Springer), 255–264. doi:10.1007/978-1-4614-1689-0_15

Lee, S., Choi, S., Kim, Y. J., Kim, B.-J., Hwang, H., Park, T., et al. (2016). Pathway-based Approach Using Hierarchical Components of Collapsed Rare Variants. *Bioinformatics* 32, i586–i594. doi:10.1093/bioinformatics/btw425

Lee, S., Choi, S., Qiao, D., Cho, M., Silverman, E. K., Park, T., et al. (2018). WISARD: Workbench for Integrated Superfast Association Studies for Related Datasets. *BMC Med. Genomics* 11, 39–44. doi:10.1186/s12920-018-0345-y

Lu, Y. C., Wang, P., Wu, Q. G., Zhang, R. K., Kong, A., Li, Y. F., et al. (2019). Hsp74/14-3-3σ Complex Mediates Centrosome Amplification by High Glucose, Insulin, and Palmitic Acid. *Proteomics* 19, 1800197. doi:10.1002/pmic.201800197

Maebashi, M., Makino, Y., Furukawa, Y., Ohinata, K., Kimura, S., and Sato, T. (1993). Therapeutic Evaluation of the Effect of Biotin on Hyperglycemia in Patients with Non-insulin Dependent Diabetes Mellitus. *J. Clin. Biochem. Nutr.* 14, 211–218. doi:10.3164/jcbn.14.211

Metzker, M. L. (2010). Sequencing Technologies - the Next Generation. *Nat. Rev. Genet.* 11, 31–46. doi:10.1038/nrg2626

Moayyeri, A., Cheung, C.-L., Tan, K. C., Morris, J. A., Cerani, A., Mohney, R. P., et al. (2018). Metabolomic Pathways to Osteoporosis in Middle-Aged Women: A Genome-Metabolome-Wide Mendelian Randomization Study. *J. Bone Miner Res.* 33, 643–650. doi:10.1002/jbmr.3358

Murea, M., Ma, L., and Freedman, B. I. (2012). Genetic and Environmental Factors Associated with Type 2 Diabetes and Diabetic Vascular Complications. *Rev. Diabet Stud.* 9, 6–22. doi:10.1900/rds.2012.9.6

Ouyang, Y., Qiu, G., Zhao, X., Su, B., Feng, D., Lv, W., et al. (2021). Metabolome-Genome-Wide Association Study (mGWAS) Reveals Novel Metabolites Associated with Future Type 2 Diabetes Risk and Susceptibility Loci in a Case-Control Study in a Chinese Prospective Cohort. *Glob. Challenges* 5, 2000088. doi:10.1002/gch2.202000088

Park, T. J., Lee, H. S., Kim, Y. J., and Kim, B. J. (2019). Identification of Novel Non-synonymous Variants Associated with Type 2 Diabetes-Related Metabolites in Korean Population. *Biosci. Rep.* 39, BSR20190078. doi:10.1042/BSR20190078

Privé, F., Vilhjálmsson, B. J., Aschard, H., and Blum, M. G. (2019). Making the Most of Clumping and Thresholding for Polygenic Scores. *Am. J. Hum. Genet.* 105, 1213–1221. doi:10.1016/j.ajhg.2019.11.001

Reddi, A., Deangelis, B., Frank, O., Lasker, N., and Baker, H. (1988). Biotin Supplementation Improves Glucose and Insulin Tolerances in Genetically Diabetic KK Mice. *Life Sci.* 42, 1323–1330. doi:10.1016/0024-3205(88)90226-3

Revilla-Monsalve, C., Zendejas-Ruiz, I., Islas-Andrade, S., Báez-Saldaña, A., Palomino-Garibay, M. A., Hernández-Quiróz, P. M., et al. (2006). Biotin Supplementation Reduces Plasma Triacylglycerol and VLDL in Type 2 Diabetic Patients and in Nondiabetic Subjects with Hypertriglyceridemia. *Biomed. Pharmacother.* 60, 182–185. doi:10.1016/j.biopha.2006.03.005

Shan, Z., Sun, T., Huang, H., Chen, S., Chen, L., Luo, C., et al. (2017). Association between Microbiota-dependent Metabolite Trimethylamine-N-Oxide and Type 2 Diabetes. *Am. J. Clin. Nutr.* 106, 888–894. doi:10.3945/ajcn.117.157107

Shin, N. R., Gu, N., Choi, H. S., and Kim, H. (2020). Combined Effects of Scutellaria Baicalensis with Metformin on Glucose Tolerance of Patients with Type 2 Diabetes via Gut Microbiota Modulation. *Am. J. Physiol.-Endocrinol. Metab.* 318, E52–E61. doi:10.1152/ajpendo.00221.2019

Suhre, K., and Gieger, C. (2012). Genetic Variation in Metabolic Phenotypes: Study Designs and Applications. *Nat. Rev. Genet.* 13, 759–769. doi:10.1038/nrg3314

Tang, W. H. W., Wang, Z., Li, X. S., Fan, Y., Li, D. S., Wu, Y., et al. (2017). Increased Trimethylamine N-Oxide Portends High Mortality Risk Independent of Glycemic Control in Patients with Type 2 Diabetes Mellitus. *Clin. Chem.* 63, 297–306. doi:10.1373/clinchem.2016.263640

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., et al. (2011). Quality Control Procedures for Genome-wide Association Studies. *Curr. Protoc. Hum. Genet.* Chapter 1, Unit1.19. doi:10.1002/0471142905.hg0119s68

Van Doorn, M., Vogels, J., Tas, A., Van Hoogdalem, E. J., Burggraaf, J., Cohen, A., et al. (2007). Evaluation of Metabolite Profiles as Biomarkers for the Pharmacological Effects of Thiazolidinediones in Type 2 Diabetes Mellitus Patients and Healthy Volunteers. *Br. J. Clin. Pharmacol.* 63, 562–574. doi:10.1111/j.1365-2125.2006.02816.x

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* 97, 576–592. doi:10.1016/j.ajhg.2015.09.001

Weyer, C., Tataranni, P. A., Bogardus, C., and Pratley, R. E. (2001). Insulin Resistance and Insulin Secretory Dysfunction Are Independent Predictors of Worsening of Glucose Tolerance during Each Stage of Type 2 Diabetes Development. *Diabetes care* 24, 89–94. doi:10.2337/diacare.24.1.89

Xie, Z., Su, W., Guo, Z., Pang, H., Post, S., and Gong, M. (2006). Up-Regulation of CPI-17 Phosphorylation in Diabetic Vasculature and High Glucose Cultured Vascular Smooth Muscle Cells. *Cardiovasc. Res.* 69, 491–501. doi:10.1016/j.cardiores.2005.11.002

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* 88, 76–82. doi:10.1016/j.ajhg.2010.11.011

Zhang, H., Osada, K., Sone, H., and Furukawa, Y. (1997). Biotin Administration Improves the Impaired Glucose Tolerance of Streptozotocin-Induced Diabetic Wistar Rats. *J. Nutr. Sci. Vitaminol.* 43, 271–280. doi:10.3177/jnsv.43.271

Zhang, G., Saito, R., and Sharma, K. (2017). A Metabolite-GWAS (mGWAS) Approach to Unveil Chronic Kidney Disease Progression. *Kidney Int.* 91, 1274–1276. doi:10.1016/j.kint.2017.03.022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 23 May 2022
doi: 10.3389/fgene.2022.769828

Multistage Combination Classifier Augmented Model for Protein Secondary Structure Prediction

Xu Zhang[1], Yiwei Liu[2], Yaming Wang[3], Liang Zhang[4], Lin Feng[2], Bo Jin[2] and Hongzhe Zhang[1]*

[1]College of Mechanical Engineering, Dalian University of Technology, Dalian, China, [2]School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian, China, [3]The First Affiliated Hospital, Dalian Medical University, Dalian, China, [4]International Business School, Dongbei University of Finance and Economics, Dalian, China

In the field of bioinformatics, understanding protein secondary structure is very important for exploring diseases and finding new treatments. Considering that the physical experiment-based protein secondary structure prediction methods are time-consuming and expensive, some pattern recognition and machine learning methods are proposed. However, most of the methods achieve quite similar performance, which seems to reach a model capacity bottleneck. As both model design and learning process can affect the model learning capacity, we pay attention to the latter part. To this end, a framework called Multistage Combination Classifier Augmented Model (MCCM) is proposed to solve the protein secondary structure prediction task. Specifically, first, a feature extraction module is introduced to extract features with different levels of learning difficulties. Second, multistage combination classifiers are proposed to learn decision boundaries for easy and hard samples, respectively, with the latter penalizing the loss value of the hard samples and finally improving the prediction performance of hard samples. Third, based on the Dirichlet distribution and information entropy measurement, a sample difficulty discrimination module is designed to assign samples with different learning difficulty levels to the aforementioned classifiers. The experimental results on the publicly available benchmark CB513 dataset show that our method outperforms most state-of-the-art models.

Keywords: genetics, biology, protein secondary structure, deep learning, combination classifier, amino acid sequence

OPEN ACCESS

Edited by:
Jiajie Peng, Northwestern Polytechnical University, China

Reviewed by:
Pengyang Wang, University of Macau, China
Ping Zhang, The Ohio State University, United States

****Correspondence:***
Bo Jin
jinbo@dlut.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *02 September 2021*
Accepted: *25 January 2022*
Published: *23 May 2022*

Citation:
Zhang X, Liu Y, Wang Y, Zhang L, Feng L, Jin B and Zhang H (2022) Multistage Combination Classifier Augmented Model for Protein Secondary Structure Prediction. Front. Genet. 13:769828. doi: 10.3389/fgene.2022.769828

INTRODUCTION

Gene controls the individual characters of biology through the guidance of protein synthesis to express its own genetic information. With the completion of the human genome project, scientists have never stopped studying the protein structure. Understanding protein secondary structure is very important for exploring diseases and finding new treatments (Huang et al., 2016; Li et al., 2021). Protein structure prediction is a very important research topic in the field of bioinformatics. Protein is the material basis of life activities, the basic organic matter of cells, and the main undertaker of life activities. Proteins can be folded into different structures or conformations, showing the feasibility of various biological processes in organisms. The protein structure determines its function, so the prediction of protein structure has great research value. In the field of bioinformatics, it is difficult to predict the spatial protein structure from the primary structure, so the prediction of the protein

secondary structure has attracted much attention (Zhang, 2008; Källberg et al., 2012). Protein secondary structures refer to the local spatial structure of the polypeptide chain skeleton, not considering the conformation of the side chain and the spatial arrangement of the whole peptide chain. Besides, protein secondary structures are stabilized by hydrogen bonds on the backbone and are considered the linkages between primary sequences and tertiary structures (Myers and Oas, 2001). According to the distinct hydrogen bonding modes, generally, three types of secondary structures have been identified, namely helix (H), strand (E), and coil (C), where the helix and strand structures are most common in nature (Pauling et al., 1951). In the new classification calculated by the DSSP algorithm, the previous three states are extended to eight states, including α-helix (H), 3_{10} helix (G), π-helix (I), β-strand (E), β-bridge (B), β-turn (T), bend (S), and coil (C) (Kabsch and Sander, 1983), among which the α-helix and β-strand are the principal structure features (Lyu et al., 2021).

In the field of genetics and bioinformatics, protein secondary structure prediction is intended to predict the three-dimensional structure of a protein from its amino acid sequence (Drozdetskiy et al., 2015). The protein structure prediction is very important for understanding the relationship between protein structure and its function. Experimental protein structure determination methods include X-ray crystallography, nuclear magnetic resonance spectroscopy, and electron microscopy. However, all of these methods are very time-consuming and expensive and require expertise. What is more, at present, the growth rate of the protein sequence is much higher than the chemical or biological protein structure determination methods (Fang et al., 2020). Hence, it is very urgent to explore the protein secondary structure prediction methods. Although the three-dimensional structure of a protein cannot be accurately predicted directly from the amino acid sequence of the protein, we can predict the protein secondary structure to understand the three-dimensional structure of the protein. The protein secondary structure reserves part of the three-dimensional structure information and can help understand the three-dimensional morphology of the amino acid in the primary structure (Hanson et al., 2019).

Due to the high application value of the protein secondary structure prediction in many biological aspects, plenty of related algorithms based on deep learning methods have been proposed over the years (Li and Yu., 2016; Wang et al., 2016; Heffernan et al., 2017; Fang et al., 2018; Zhang et al., 2018; Uddin et al., 2020; Guo et al., 2021; Lyu et al., 2021; Drori et al., 2018). Current methods mainly utilize the convolutional and recurrent neural network to extract different protein features and then apply them to protein secondary structure prediction. For example, Li and Yu (2016) proposed an end-to-end deep network to predict the secondary structure of proteins from the integrated local and global context features, which leveraged convolutional neural networks with different kernel sizes to extract multiscale local contextual features and a bidirectional neural network consisting of the gated recurrent unit to capture global contextual features. Wang et al. (2016) presented Deep Convolutional Neural Fields (DeepCNF) for protein secondary structure prediction, which can model not only complex sequence-structure relationship by a deep hierarchical architecture but also interdependency between adjacent protein secondary labels, so it is much more powerful than traditional Convolutional Neural Fields. Lyu et al. (2021) presented a reductive deep learning model MLPRNN to predict either 3-state or 8-state protein secondary structures. Besides, Uddin et al. (2020) incorporated a self-attention mechanism within the Deep Inception-Inside-Inception network (Fang et al., 2018) to capture both the short- and long-range interactions among the amino acid residues. Guo et al. (2021). Integrated and developed multiple advanced deep learning architectures (DNSS2) to further improve secondary structure prediction. As described above, most researchers currently focus on exploring the complex deep learning models, and a few try to solve the protein secondary structure prediction task from the perspective of model learning or training methods, for example, "ELF: An Early-Exiting Framework for Long-Tailed Classification" (Duggal et al., 2020).

Real data usually follow a long-tailed distribution, most concentrated in only a few classes. On datasets following this distribution, neural networks usually cannot deal well with all classes (majority or rareness classes). If the model performs well on majority classes, it tends to perform poorly on the rareness classes and vice versa, resulting in poor performance. The protein secondary structure prediction task also shows a similar problem. For example, we visualize the CB6133-filtered and CB513 datasets (Zhou and Troyanskaya, 2014) and find an imbalance problem in the protein secondary structure labels distribution. For example, the number of labels α-helix (H), β-strand (E), and coil (C) is much greater than other labels. This imbalance problem has traditionally been solved by resampling the data (e.g., undersampling and oversampling) (Chawla et al., 2002; Minlong et al., 2019) or reshaping the loss function (e.g., loss reweighting and regularization) (Cao et al., 2019; Cui et al., 2019). However, by treating each example within a class equally, these methods fail to account for the important notion of example hardness. In other words, within each class, some examples are easier to classify than others (Duggal et al., 2020). Hence, "ELF: An Early-Exiting Framework for Long-Tailed Classification" is proposed to overcome the above-described limitations and address the challenge of data imbalance. ELF incorporates the notion of hardness into the learning process and can induce the neural network to increasingly focus on hard examples because they contribute more to the overall network loss. Hence, it frees up additional model capacity to distinguish difficult examples and can improve the classification performance of the model.

To our knowledge, few studies try to solve the protein secondary structure prediction task from the perspective of the model learning process. This study proposes a framework called Multistage Combination Classifier Augmented Model (MCCM) to solve that task and fill in the blanks. We first introduce a feature extraction module to extract features with different learning difficulty levels. Then, we design two classifiers that can learn the decision boundaries for easy and hard samples, respectively. Finally, we propose a sample learning difficulty discrimination module *via* exploring two strategies. Specifically, the first strategy is label-dependent, assuming the sample is hard if it is misclassified. However, the actual data is lack of labels. Hence,

the second strategy utilizes Dirichlet distribution and information entropy measurement. The experimental results based on the first method and the benchmark CB513 dataset show that our proposed framework outperforms other state-of-the-art models by a large margin, indicating that if the multilevel samples discriminating module can be designed effectively, our framework can obtain state-of-the-art performance. Furthermore, the results based on the second method also show that our model outperforms most state-of-the-art models. In this work, we made the following key contributions:

- We are first to develop a Multistage Combination Classifier Augmented Framework for protein secondary structure prediction task. It consists of multilevel (easy or hard level in this study) features extraction, multistage combination classifiers, and multilevel samples discrimination module. The last module is realized based on label-dependent and label-independent methods, respectively.
- For our core multilevel samples discrimination module, a label-independent measurement standard to discriminate the easy and hard samples is first explored by our work based on Dirichlet distribution and information entropy theory. The Dirichlet distribution is designed to measure the model confidence based on subjective logic theory. The information entropy is designed to evaluate whether the Dirichlet distribution shows a highly confident distribution and, thus, capture the easy samples that tend to be classified accurately by the easy classifier.
- The results based on the label-independent method show that our model outperforms most state-of-the-art methods, indicating that the designed multilevel samples discrimination module herein is effective. The excellent result based on the label-dependent method means that our framework can obtain a state-of-the-art performance if the multilevel samples discriminating module is designed appropriately. Hence, our work not only offers a new idea to deal with the protein secondary structure prediction task but also leaves room for further research focusing on how to design a more effective multilevel samples discrimination module.

METHODS AND MATERIALS

Benchmark Datasets

In the field of protein secondary structure prediction in genetics and bioinformatics, CB6133-filtered and CB513 datasets (Zhou and Troyanskaya, 2014) are two benchmark datasets widely used by the researchers (Li and Yu., 2016; Fang et al., 2018; Zhang et al., 2018; Guo et al., 2021; Lyu et al., 2021). The CB6133-filtered dataset is filtered to remove redundancy with the CB513 dataset (for testing performance on the CB513 dataset). In particular, the CB6133-filtered dataset is used to train the model, and the CB513 dataset is used to test the model. The training CB6133-filtered dataset is a large non-homologous sequence and structure containing 5,600 training sequences. The dataset is made by the PISCES Cull PDB server, a public server for screening protein sequence sets from the Protein Data Bank (PDB) according to sequence identification and structural quality standards (Wang and Dunbrick, 2003). The testing dataset CB513 was introduced by Cuff and Barton (Cuff and Barton, 1999, 2000), which contains 514 sequences. The two available benchmark datasets can be obtained by Zhou's website.

Input Features

Considering the difficulty of protein secondary structure prediction in genetics and bioinformatics, we use four types of features to characterize each residue in a protein sequence, including 21-dimensional amino acid residues, one-hot coding, and the sequence of 21-dimensional profile features, obtained from the PSI-BLAST (Altschul et al., 1997) log file and rescaled by a logistic function (Jones, 1999). Furthermore, the seven-dimensional physical property features (Jens et al., 2001) were previously used for the protein structure and property prediction by researchers (Heffernan et al., 2017) and obtained a good performance. The physical properties include steric parameters (graph-shape index), polarizability, normalized van der Waals volume (VDWV), hydrophobicity, isoelectric point, helix probability, and sheet probability. We also take them as one of the input features, and the features can be downloaded from Meiler's study (Jens et al., 2001). Finally, the one-dimensional conservation score was obtained by applying the method (Quan et al., 2016):

$$R = log\,20 + \sum_{1}^{20} L_i \, log\, L_i. \tag{1}$$

The residues are transformed according to the frequency distribution of amino acids in the corresponding column of homologous protein multiple sequence alignment. The score information in the profile features was calculated from this probability. Residue score in the *i*th column was calculated as follows (Altschul et al., 1997):

$$S_i = \left[\ln\left(\frac{L_i}{P_i}\right) \right] \Big/ \lambda_u, \tag{2}$$

where L_i is the predicted probability that a properly aligned homologous protein has amino acid i in that column and P_i is the background probability. λ_u is 0.3176. L_i is defined as

$$L_i = \exp\left(S_i \cdot \lambda_u\right) \cdot P_i. \tag{3}$$

Model Design

The proposed model for protein secondary structure prediction in genetics and bioinformatics consists of a feature extracting module and a Multistage Combination Classifier Module. This section firstly introduces the two modules separately and then explains the overall architecture in detail.

Multilevel Features Extraction

We use a multilevel features extraction module to extract easy (low level) and hard (high level) features. The easy feature is

obtained through four linear layers and multiscale one-dimensional convolution layers. At first, we apply the four linear layers to the amino acid residues one-hot coding, sequence profile, physical property, and conservation score features, respectively. Further, we apply the concatenation function for the outputs, intended to obtain the feature representations with denser and more information. We define the concatenated outputs as

$$l = \left[l_1^1, \ldots, l_T^1, l_1^2, \ldots, l_T^2, l_1^k, \ldots, l_T^k\right], \tag{4}$$

where l_T^k denotes the output of the linear layer k and T is the index of amino acid sequence. To model local dependencies of adjacent amino acids, we leverage multiscale one-dimensional convolution layers to extract local contexts (Li and Yu., 2016):

$$c_i = F \cdot l_{i:\, i+f-1}^k = \mathrm{Relu}\left(\mathrm{w} \cdot l_{i:\, i+f-1}^k + b\right), \tag{5}$$

where $F \in \mathbb{R}^{f \cdot m}$ is a convolutional kernel, f is kernel size, and m is the feature dimensionality of the concatenated outputs of the four linear layers l. w and b is the trainable parameters of the convolution layers. In this study, f of the three one-dimensional convolution layers is 5, 9, and 13, respectively. We define the concatenated outputs of the multiscale one-dimensional convolution layers as

$$c = \left[c_1^1, \ldots, c_T^1, cl_1^2, \ldots, c_T^2, cl_1^k, \ldots, c_T^k\right], \tag{6}$$

where c_T^k denotes the output of the convolution layer k.

Then, based on the obtained easy feature c, the hard feature is further extracted by one Gate Recurrent Unit ($gru(\cdot)$) and the attention mechanism. $gru(\cdot)$ is designed to capture the global contexts in the amino acid sequences. Defining the input of the GRU as (c_t^k, h_{t-1}), the mechanism of a GRU can be presented as follows:

$$r_t = sigm(W_{cr} \cdot c_t + W_{hr} \cdot h_{t-1} + b_r), \tag{7}$$

$$u_t = sigm(W_{cu} \cdot c_t + W_{hu} \cdot h_{t-1} + b_u), \tag{8}$$

$$\tilde{h}_t = tanh\left(W_{c\tilde{h}} \cdot c_t + W_{h\tilde{h}} \cdot \left(r_t \odot h_{t-1} + b_{\tilde{h}}\right)\right), \tag{9}$$

$$h_t = u_t \odot h_{t-1} + (1 - u_t) \odot \tilde{h}_t, \tag{10}$$

where r_t is the activation of the reset gate; u_t is the activation of the update gate; $\tilde{h}_t$ is the internal memory cell; h_t is the GRU output; W_{cr}, W_{cr}, W_{cu}, W_{hu}, $W_{c\tilde{h}}$, and $W_{h\tilde{h}}$ are weight matrices; and b_r, b_u, and $b_{\tilde{h}}$ are bias terms. Besides, $\odot$, $sigm$, and $tanh$ denote element-wise multiplication, sigmoid, and hyperbolic functions, respectively. Further, the sequential attention mechanism (SAM) has been widely used in the LSTM-based solutions for sequential learning problems (Feng et al., 2019). In this study, considering the global contexts h_t of different amino acid sequence steps could contribute differently to the representation of the whole amino acid sequences. We use the attention mechanism to compress the hidden representations of global contexts h_t at different sequence steps into an overall representation with adaptive weights:

$$\tilde{\alpha}_t = s_a^T \, tanh(W_a \cdot h_t + b_a), \tag{11}$$

$$\alpha_t = \frac{exp^{\tilde{\alpha}_t}}{\sum_{t=1}^{T} exp^{\tilde{\alpha}_t}}, \tag{12}$$

$$\alpha_h = \sum_{t=1}^{T} \alpha_t h_t, \tag{13}$$

where W_a, s_a, and b_a are trainable parameters and α_h denotes the important contexts information, aggregating from the global contexts h_t. Although α_h aggregated most part of the important contexts' information, it also may lead to losing part of important information more or less. Hence, we apply the concatenation function to the original global contexts h_t and the aggregated contexts information α_h. At last, the obtained local contexts, global contexts, and aggregated global contexts through SAM are concatenated together as the hard features:

$$v = [c, h, \alpha_h]. \tag{14}$$

Finally, the easy feature $\boldsymbol{c}$ is sent to the easy classifier, and the hard feature $\boldsymbol{v}$ is sent to the hard classifier.

Multistage Combination Classifier Module

Predicting protein secondary structure in genetics and bioinformatics is a challenging task that we try to solve from the perspective of the model learning method. On the one hand, the existing research results point out that, within different classes of all samples (the classes are either majority or minority), some examples are easier than the others (Duggal et al., 2020). On the other hand, different people may be suitable for different work. Similarly, the different classifiers may be suitable for classifying different samples. Following the theory and intuition, we design two classifier branches in the model to deal with samples with different difficulty levels. The first classifier branch comprises a simple linear layer, which aims to deal with the simple samples (easy to classify). The second classifier branch comprises a multi-layer perceptron ($MLP(\cdot)$), which is more complex than the first classifier and aims to deal with the hard samples (hard to classify). The first classifier branch can correctly classify some easy samples and serve as a filter to filter them out, avoiding being sent to the second complex classifier. We regard the remaining samples, classified by the first classifier incorrectly, as hard samples and further send them to the second complex classifier. After the computation of each classifier, we calculate the cross-entropy loss between the predicted outputs and ground truth labels:

$$\mathcal{L}_{\mathrm{C}} = \frac{1}{B} \sum_{i}^{B} - \sum_{j=1}^{Z} y_{ij} \left(\log\left(p_{ij}\right)\right), \tag{15}$$

where B is the number of batch samples and Z is the number of target labels. We can further obtain the cross-entropy loss computed based on the easy and hard classifier and describe them as $\mathcal{L}_{\mathrm{C\_easy}}$ and $\mathcal{L}_{\mathrm{C\_hard}}$, respectively. After the computation of the first simple classifier, we can obtain the loss value of all samples. Further, we can obtain the loss value of the hard samples after the computation of the second hard classifier. Hence, the loss value of the harder samples is increased in general, and the model is induced to pay more attention to harder samples and improve the classification performance. The final loss function is the sum of the cross-entropy loss of the easy and hard versions:

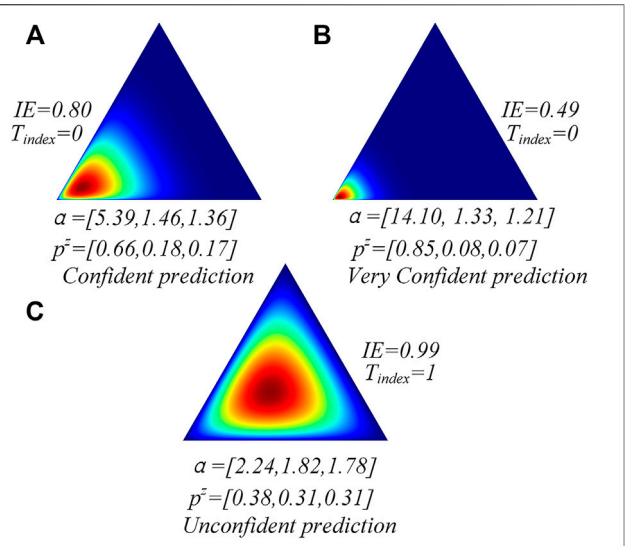

FIGURE 1 | Prediction of the Dirichlet distribution of the three amino acid samples' analysis.

$$\mathcal{L}_{\text{C\_all}} = \mathcal{L}_{\text{C\_easy}} + \mathcal{L}_{\text{C\_hard}}. \tag{16}$$

Sample Difficulty Discrimination Module

We have designed the easy and hard classifiers to deal with different samples. However, we need to further design a measurement standard to discriminate between the easy and hard samples among all samples. For the model, a label is an ideal tool to realize our purpose. If samples are classified accurately according to their labels, we regard them as easy samples and the others as hard samples sent to the hard classifier. In this way, the different classifiers can be assigned suitable samples, and our model can be trained well. However, the actual data is lack of labels. Hence, we just can design the measurement standard to get close to the ideal effect in all possible ways. In this study, we design a measurement standard based on subjective logic (SL) and Dirichlet distribution with Z parameters $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_Z]$, and α are called subjective opinions (Dempster, 2008; Josang, 2016). If the model has a highly motivated subjective opinion on one class of one test sample, it means that the model is confident to classify it accurately after being trained on the training data. For example, as shown in **Figure 1**, the easy classifier classifies the amino acid into three states (H, E, and C). IE denotes the information entropy, T_{index} denotes the index of True label, and p^z denotes the expectation of the Dirichlet distribution. We will further discuss them in the following part. Based on the subjective logic theory, we know that if the predicted α of certain amino acids are [14.10, 1.33, and 1.21] (each α corresponds to one state), the easy classifier is very confident to classify the amino accurately. Hence, following a Dirichlet distribution, the subjective multinomial opinion will yield a sharp distribution on one corner of the simplex (**Figures 1A, B**). However, if the predicted α are [2.24, 1.82, and 1.78], as shown in **Figure 1C**, the model is not confident to classify it accurately, and it should be sent to a hard classifier. In this condition, the multinomial opinion will yield a central distribution (**Figure 1B**). The Dirichlet distribution is a probability density function (pdf) for possible values of the probability mass function (pmf) p and can be expressed by Z parameters α:

$$Dir(p|\alpha) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{Z} p_i^{\alpha_{i-1}}, \\ 0 \quad otherwise \end{cases}, \tag{17}$$

where p is the probability mass function and $\alpha = [\alpha_1, \ldots, \alpha_Z]$ are the parameters of Dirichlet distribution. Z denotes the label category. $B(\alpha)$ is a polynomial beta function in Z dimension [36]. Based on SL, the expectation of Dirichlet distribution based on neural network evidence theory can be computed as follows:

$$p^z = \alpha_z / S, \tag{18}$$

$$S = \sum_{z=1}^{Z} \alpha_z, \tag{19}$$

$$\alpha_z = e_z + 1, \tag{20}$$

$$e_z = \zeta(\hat{y}^s), \tag{21}$$

where α_z are Dirichlet parameters, $\hat{y}^s$ is the output vector before being sent to the *softmax* layer, and $\zeta(\cdot)$ denotes an activation layer (e.g., ReLU). e_z is the amount of evidence and S is the Dirichlet strength. By minimizing the mean square error loss based on the Dirichlet parameters, the Dirichlet distribution can be optimized according to the loss function as

$$L^{dir} = \sum_{b=1}^{B} (y_b - p_b)^2 + \frac{p_b(1 - p_b)}{(S_b + 1)}, \tag{22}$$

where B is the batch size of the samples, y_b is the real label of a single sample, p_b is the Dirichlet distribution expectation of a single sample, and S_b is the Dirichlet strength of a single sample.

Finally, we use the information entropy (IE) to know whether the easy classifier has a highly motivated subjective opinion on the samples. Given the predicted Dirichlet distribution parameters $[\alpha_1, \alpha_2, \ldots, \alpha_Z]$, we can compute p^z. Further, we can compute the information entropy of p^z, which is defined as

$$H(p^z) = -\sum p(p^z) log_2(p^z). \tag{23}$$

As shown in **Figure 1**, if the easy classifier is very confident to classify the sample accurately, its information entropy tends to be lower than other conditions. We can also find that the classifier with low information entropy shows a highly motivated subjective opinion on the current samples and classifies them accurately (**Figures 1A,B**). However, the classifier with high information entropy shows a uniform subjective opinion on the current sample labels and classifies it incorrectly. Hence, the information entropy can be used to help the model discriminate between the easy and hard samples. We define the discriminating process as

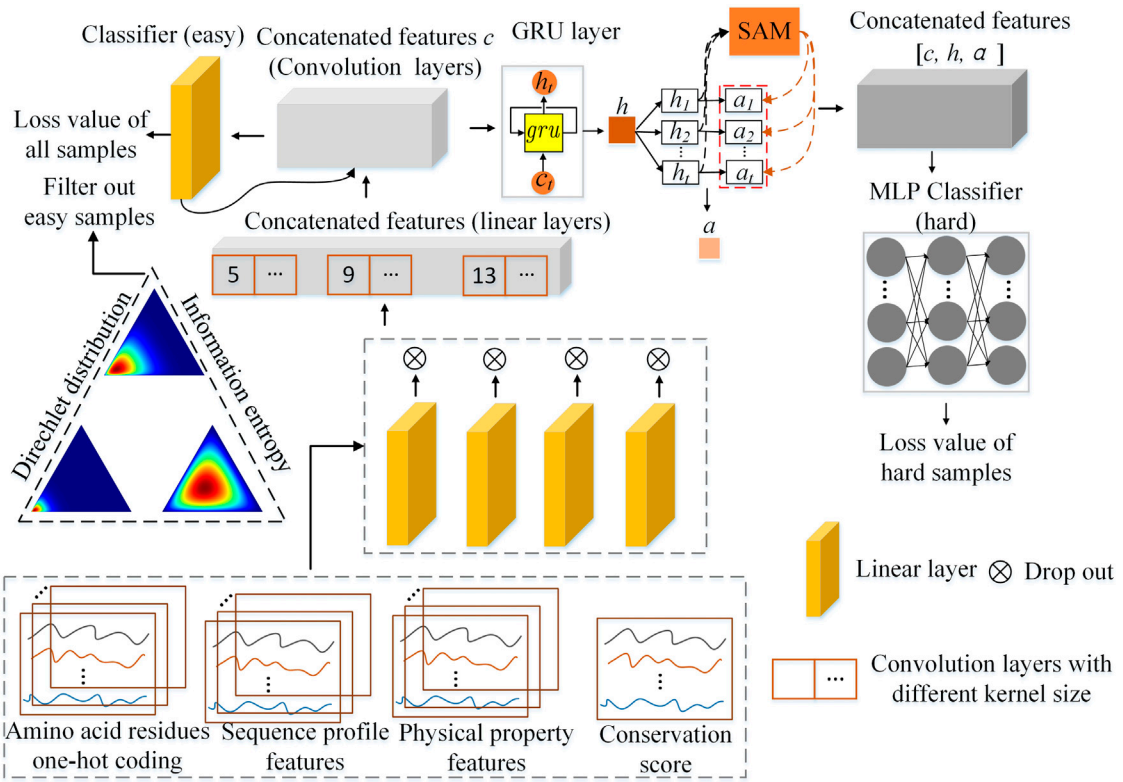

FIGURE 2 | The overall architecture of the Multistage Combination Classifier Augmented Model (MCCM) for protein secondary structure prediction in genetics and bioinformatics.

$$samples_{cur} = \begin{cases} samples_{easy}, & if\ argmax(p_b) = y_b\ and\ H(p^z) < H(p^z)_{per} \\ samples_{hard}, & otherwise \end{cases}, \tag{24}$$

where $H(p^z)_{per}$ is *per* th percentile of $H(p^z)$, which is a threshold to distinguish the samples into hard or easy samples. In particular, in the training process, if the samples are classified correctly and their information entropy is lower than the threshold, they will not be sent to the hard classifier. Otherwise, the samples will be sent to the hard classifier again. In the test process, there is no need to know whether the samples are classified correctly, and the easy or hard samples are only divided based on the information entropy. The samples with the high information entropy will be sent to the hard classifier for the final prediction result.

Hence, the final loss function of our model can be obtained by uniting **Eqs. 16**, **22**:

$$\mathcal{L}_{\text{C\_final}} = \mathcal{L}_{\text{C\_easy}} + \beta \cdot L_{easy}^{dir} + \mathcal{L}_{\text{C\_hard}} + \beta \cdot L_{hard}^{dir}. \tag{25}$$

According to **Eq. 22**, L_{easy}^{dir} and L_{hard}^{dir} are calculated based on the output of the easy and hard classifiers, respectively.

The architecture of the Multistage Combination Classifier Augmented Model (MCCM) is shown in **Figure 2**. After preprocessing the dataset, 50-dimensional features are obtained and taken as the input features, including the 21-dimensional amino acid residues one-hot coding, 21-dimensional sequence profile, 7-dimensional physical property, and 1-dimensional conservation score. The features are first preprocessed into easy ones through four linear layers and multiscale one-dimensional convolution layers. Based on the Dirichlet distribution and information entropy, the samples are divided into easy and hard ones by an easy classifier (a simple linear layer). Then, the easy feature is further preprocessed into a hard one through ***gru***$(\cdot)$ and the attention mechanism SAM. Finally, the hard samples are sent into a hard classifier (***MLP***$(\cdot)$).

Implementation Details

The hidden sizes of the four linear layers used for the 21-dimensional amino acid residues one-hot coding, 21-dimensional sequence profile, 7-dimensional physical property, and 1-dimensional conservation score features are 64, 128, 32, and 16, respectively. The hidden size of the multiscale one-dimensional convolution layers is 64 and the corresponding kernel sizes are 5, 9, and 13, respectively. The GRU layer is bidirectional and the hidden size is 256. The hidden size of the linear layer used in the attention mechanism is 256. The hidden sizes of the first two layers used in MLP are 512 and 1,024. The models are optimized by Adam optimizer, and the learning rates are set to 0.0005. During training, the dropout function can

TABLE 1 | Q3 and Q8 accuracy of different algorithms on the public CB513 dataset.

Algorithms	Q3	Q8
DeepCNF (2016)	81.80	69.1
DCRNN (2018)	-	69.70
eCRRNN (2018)	81.20	70.2
DNSS2 (2021)	82.56	73.36
BLSTM (2015)	-	67.40
GSN (2014)	-	66.40
SSpro, free (2014)	78.50	63.50
JPRED4 (2015)	81.70	-
SecNet (2020)	84.30	72.30
$MCCM_{dir}$	82.12	69.79
$MCCM_{easy}$	86.94	71.78
MCCM	**96.31**	**83.74**

The bold values denote the best values of performance metrics.

randomly zero some of the elements of the input tensor with probability τ using samples from a Bernoulli distribution. Herein, the dropout function is used in four linear layers and the MLP layers and τ is set as 0.5. The percentile of $H(p^z)$ used in this study is 15, 30, and 35. The β used in **Eq. 22** is 1. All results have been produced based on the same hardware environment: Intel (R) Core (TM) CPU I7-10700 @ 2.90 GHz 16 cores. Finally, we define the proposed label-dependent model (only used to explore the theoretical best performance) equipped with both $\mathcal{L}_{C\_easy}$ and $\mathcal{L}_{C\_hard}$ as MCCM. The proposed label-dependent model equipped with only $\mathcal{L}_{C\_easy}$ is defined as $MCCM_{easy}$, which means there is no backpropagation operation through the loss function $\mathcal{L}_{C\_hard}$. The proposed label-independent model (use the evidence and information entropy theory to divide the samples into easy and hard ones) is $MCCM_{dir}$.

Performance Evaluation

In the field of protein secondary structure prediction in genetics and bioinformatics, the Q score measurement formulated as **Eq. 6** has been widely used to evaluate the performance of the proposed models. It measures the percentage of residues for which the predicted secondary structures are correct (Wang et al., 2016):

$$Q_k = 100\% \times \frac{\sum_{i=1}^{k} N_{correct}(i)}{N}, \tag{26}$$

where k indicates the number of classes, for example, Q_3 score (k = 3) or Q_8 score (k = 8). Q_8 classes include α-helix (H), 3_{10} helix (G), π-helix (I), β-strand (E), β-bridge (B), β-turn (T), bend (S), and coil (C). Q_8 is transformed to Q_3 by treating the label (B, E) as E, (G, I, H) as H, and (S, T, C) as C.

RESULTS AND DISCUSSION

Experimental Results of Evaluating Indicators

The evaluation results of the proposed model based on the public CB513 test dataset are shown in **Table 1**. $MCCM_{dir}$ means using the Dirichlet distribution and information entropy to divide the samples into easy and hard ones. In

TABLE 2 | Q3 and Q8 accuracy of variant models on the public CB513 dataset.

Algorithms	Q3	Q8
$MCCM_{c1}$	79.93	66.30
$MCCM_{c2}$	81.45	68.92
$MCCM_{conf}$	81.00	66.42
$MCCM_{dir}$	**82.12**	**69.79**

The bold values denote the best values of performance metrics.

$MCCM_{easy}$ and MCCM, we use label information to divide the samples, an ideal measurement method that can help us explore the theoretical best performance. Besides, $MCCM_{easy}$ means there is no backpropagation operation through the loss function $\mathcal{L}_{C\_hard}$ and the model will not be induced to pay more attention to hard samples. The results of the benchmark methods on CB513 datasets are obtained from (Shapovalov et al., 2020), except for the DNSS2, which is obtained from Guo et al. (2021). We can find that the Q3 and Q8 accuracy of the $MCCM_{dir}$ is better than most of the benchmark methods, denoting that the designed method is based on Dirichlet distribution and information entropy to distinguish the hard or easy samples is effective. Besides, note that, due to our computer resource constraints, there are only two designed classifiers and corresponding feature extractors in our framework, which limits the performance of our model. Moreover, compared with $MCCM_{easy}$, we can find that MCCM outperforms state-of-the-art models by a large margin in both Q3 and Q8 accuracies, which means that the model is induced to pay more attention to hard samples and improves the classification performance of the model overall through the backpropagation operation of both $\mathcal{L}_{C\_easy}$ and $\mathcal{L}_{C\_hard}$. It shows that it is reasonable to use different classifiers to classify samples with different difficulty levels, thus increasing the loss values of hard samples, inducing the model to pay more attention to hard samples. However, the label information is lacking in actual data, so the excellent performance of MCCM only denotes that if the multilevel samples discriminating module (divide samples into hard and easy ones) is designed very effectively, our method can obtain the state-of-the-art performance.

Ablation Study

This section gives a more comprehensive analysis regarding the effectiveness of the proposed framework. The different levels of classifiers equipped with the multilevel samples discriminating module can improve the performance of the prediction task. The compared variants are as follows:

$MCCM_{c1}$: the variant is the front part of our proposed model (without the multilevel classifiers and sample difficulty discrimination module), which only uses the concatenated features c and easy classifier (see **Figure 2**) to deal with the prediction task. This model only uses the convolution layer to extract the features and then conducts the classification task based on the low-level features.

$MCCM_{c2}$: the variant is the part of our proposed model (without the multilevel classifiers and sample difficulty discrimination module) that uses the concatenated features

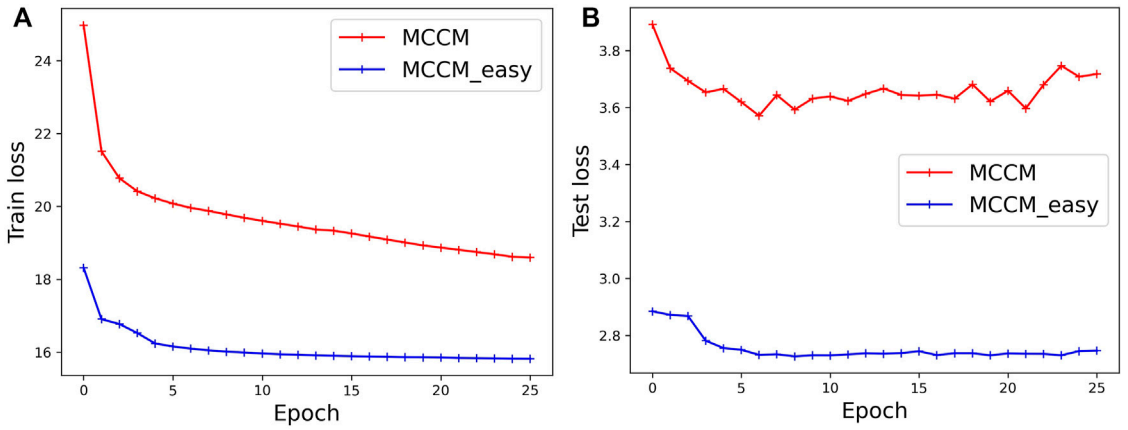

FIGURE 3 | The loss value computed on the training and test datasets.

TABLE 3 | Prediction accuracy of each label in the Q8 states based on the CB513 dataset.

Label	Types	Frequency	$MCCM_{dir}$	$MCCM_{easy}$	MCCM
H	α-Helix	30.86	91.97	89.91	**96.30**
E	β-Strand	21.25	83.67	80.08	**92.30**
C	Coil	21.14	63.73	63.92	**88.45**
T	β-Turn	11.81	53.96	**88.46**	74.37
S	Bend	9.81	26.35	23.68	**51.91**
G	3_{10} Helix	3.69	30.62	18.3	**46.87**
B	β-Bridge	1.39	4.57	3.47	**6.94**
I	π-Helix	0.04	0.00	**3.33**	**0.00**

The bold values denote the best values of performance metrics.

TABLE 4 | Q_3 confusion matrix, of 84,765 test labels ($MCCM_{dir}$, $MCCM_{easy}$, and MCCM).

Accuracy ($MCCM_{dir}$)		**Pred freq.**	**True label**		
82.12			**C**	**E**	**H**
True freq.		100%	42.76	22.65	34.59
Predicted label	C	44.53	**35.15**	3.76	3.85
	E	21.07	5.27	**16.90**	0.48
	H	34.40	4.11	0.41	**30.07**
Accuracy ($MCCM_{easy}$)		**Pred freq.**	**True label**		
86.94			**C**	**E**	**H**
True freq.		100%	42.76	22.65	34.59
Predicted label	C	36.75	**36.1**	4.87	1.79
	E	27.66	0.38	**19.65**	2.61
	H	35.59	0.26	3.14	**31.19**
Accuracy (MCCM)		**Pred freq.**	**True label**		
96.31			**C**	**E**	**H**
True freq.		100%	42.76	22.65	34.59
Predicted label	C	42.41	**41.75**	0.29	0.72
	E	22.59	0.47	**21.22**	0.95
	H	35.00	0.19	1.07	**33.33**

The bold values denote the best values of performance metrics.

[*c,h,a*] and hard classifier (see **Figure 2**) to deal with the prediction task. This model uses not only the convolution layer but also $gru(\cdot)$ and the attention mechanism to extract the features. Then, it conducts the classification task based on both the low- and high-level features.

$MCCM_{conf}$: the variant is designed based on the ELF (Duggal et al., 2020), which uses the classifier confidence to distinguish the samples into easy and hard ones. Particularly, in the training process, if the samples are classified correctly and their classifier confidence is lower than the threshold (0.9 used in ELF and this study), they will not be sent to the hard classifier. Otherwise, the samples will be sent to the hard classifier again. In the test process, the samples with low classifier confidence will be sent to the hard classifier for the final prediction result.

The experiment results are shown in **Table 2**. The performance of $MCCM_{c1}$ is the worst, and the performance of $MCCM_{c2}$ is better than it, which means that the addition of the high-level features extractor GRU and attention mechanism is effective. The performance of $MCCM_{dir}$ is better than $MCCM_{c2}$, which means that our proposed framework is effective. The designed multilevel classifiers and sample difficulty discrimination module can help the model pay more attention to the hard samples and improve the model performance. Note that, if we can increase the depth of our network and use more classifier branches, the model performance will be better. Moreover, the performance of $MCCM_{dir}$ is better than that of $MCCM_{conf}$, which means that our designed sample difficulty discrimination module is better than that proposed in ELF (Duggal et al., 2020). The Dirichlet distribution united with the information entropy can divide the samples into hard and easy ones, which is better than using the simple classifier confidience.

Analysis of the Training Process

The training loss computed on the CB6133-filtered dataset and the test loss computed on the CB513 dataset are shown in **Figure 3**. Label-dependent MCCM and label-independent

TABLE 5 | Q_8 confusion matrix of 84,765 test labels ($MCCM_{dir}$, $MCCM_{easy}$, and MCCM).

Accuracy ($MCCM_{dir}$)		**Pred freq.**	**True label**							
69.79			**C**	**B**	**E**	**G**	**I**	**H**	**S**	**T**
True freq.		100%	21.14	1.39	21.25	3.69	0.04	30.86	9.81	11.81
Predicted label	C	23.33	**13.47**	0.04	3.58	0.27	0.00	1.04	1.28	1.46
	B	0.14	0.69	**0.06**	0.33	0.02	0.00	0.11	0.09	0.09
	E	23.81	2.29	0.02	**17.78**	0.07	0.00	0.42	0.31	0.37
	G	2.57	0.65	0.00	0.21	**1.13**	0.00	0.91	0.12	0.68
	I	0.00	0.00	0.00	0.00	0.00	**0.00**	0.03	0.00	0.00
	H	33.43	0.86	0.01	0.29	0.4	0.00	**28.38**	0.1	0.82
	S	5.12	3.48	0.01	1.07	0.18	0.00	0.69	**2.58**	1.79
	T	11.6	1.9	0.	0.55	0.5	0.00	1.86	0.63	**6.37**
Accuracy ($MCCM_{easy}$)		**Pred freq.**	**True label**							
71.78			**C**	**B**	**E**	**G**	**I**	**H**	**S**	**T**
True freq.		100%	21.14	1.39	21.25	3.69	0.04	30.86	9.81	11.81
Predicted label	C	13.51	**13.51**	0.21	0.54	0.03	0.46	0.32	0.54	5.52
	B	1.14	0.00	**0.05**	0.14	0.01	0.09	0.04	0.1	0.98
	E	18.58	0.00	0.15	**17.02**	0.01	0.35	0.17	0.33	3.22
	G	0.78	0.00	0.13	0.09	**0.68**	0.11	0.11	0.2	2.38
	I	2.12	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.03
	H	28.84	0.00	0.13	0.1	0.01	0.21	**27.74**	0.21	2.46
	S	4.12	0.00	0.24	0.47	0.02	0.6	0.28	**2.32**	5.88
	T	30.91	0.00	0.24	0.22	0.02	0.3	0.17	0.41	**10.45**
Accuracy (MCCM)		**Pred freq.**	**True label**							
83.74			**C**	**B**	**E**	**G**	**I**	**H**	**S**	**T**
True freq.		100%	21.14	1.39	21.25	3.69	0.04	30.86	9.81	11.81
Predicted label	C	23.31	**18.7**	0.02	0.45	0.31	0.00	0.38	0.56	0.71
	B	0.21	0.48	**0.10**	0.29	0.04	0.00	0.09	0.26	0.14
	E	22.1	0.48	0.03	**19.62**	0.11	0.00	0.1	0.63	0.30
	G	3.23	0.53	0.00	0.21	**1.73**	0.00	0.31	0.29	0.60
	I	0.00	0.02	0.00	0.00	0.01	**0.00**	0.00	0.00	0.01
	H	31.50	0.35	0.01	0.11	0.2	0.00	**29.72**	0.23	0.24
	S	7.76	1.8	0.05	0.96	0.33	0.00	0.48	**5.09**	1.10
	T	11.88	0.96	0.00	0.46	0.49	0.00	0.42	0.69	**8.79**

The bold values denote the best values of performance metrics.

$MCCM_{dir}$ are all optimized by the loss functions $\mathcal{L}_{C\_easy}$ and $\mathcal{L}_{C\_hard}$, but $MCCM_{easy}$ is only optimized by the loss function $\mathcal{L}_{C\_easy}$. Hence, the loss value of the MCCM is always greater than that of $MCCM_{easy}$. $\mathcal{L}_{C\_hard}$ is computed based on the hard samples and can induce the model to pay more attention to the hard samples (samples tend to be classified incorrectly in both majority or rareness classes).

Analysis of the Prediction of Q8 States and Confusion Matrix

In the field of protein secondary structure prediction in genetics and bioinformatics, the predictive precision for each class of Q8 would provide more useful information, and we compute the prediction accuracy of each label in the Q8 states based on the CB513 dataset. At the same time, we compute the confusion matrix to further explore the model performance. **Table 3** shows the prediction accuracy ($MCCM_{dir}$, $MCCM_{easy}$, and MCCM) of each label in the Q8 states. **Tables 4**, **5** show the Q3 and Q8 prediction confusion matrix, respectively. Associating the three tables, we can find that although $MCCM_{easy}$ is only optimized by the loss function $\mathcal{L}_{C\_easy}$, it also outperforms $MCCM_{dir}$. The main reason for the better performance is that $MCCM_{easy}$ uses the label information to divide the samples into hard and easy ones, denoting that the method to distinguish the samples is very important.

The existing research results point out that, within different classes of all samples (either the classes is majority or minority), some examples are easier than others (Duggal et al., 2020). Comparing the performance of $MCCM_{easy}$ and MCCM, we can find that the former based on loss function $\mathcal{L}_{C\_easy}$ can correctly classify the easier examples. Further, the latter, based on loss functions $\mathcal{L}_{C\_easy}$ and $\mathcal{L}_{C\_hard}$, can correctly classify the easier examples and correctly classify the remaining harder examples. For example, from **Table 3**, we can find that $MCCM_{easy}$ (optimized only by loss function $\mathcal{L}_{C\_easy}$) prediction accuracy of labels H, E, C, S, G, and B is much lower than that of MCCM (optimized by loss functions $\mathcal{L}_{C\_easy}$ and $\mathcal{L}_{C\_hard}$), which means the harder examples are

further correctly classified. However, we also can find that the MCCM prediction accuracy of labels T and I is lower than that of $MCCM_{easy}$, which may cause by the limited classifiers. On the whole, MCCM outperforms $MCCM_{easy}$ by a large margin because MCCM is optimized by the loss functions $\mathcal{L}_{C\_easy}$ and $\mathcal{L}_{C\_hard}$, increasing the loss value of the harder samples and inducing the model to pay more attention to them. Finally, the overall classification performance of the model is improved. Hence, the most harder samples (tend to be classified incorrectly by $MCCM_{easy}$) are further classified correctly by MCCM, which can be known in the shown tables.

Associating the performance of $MCCM_{dir}$ and MCCM, we can induce that if we can design a method to distinguish the samples well (the discrimination effect is close to using label information), our method can obtain the state-of-the-art performance. Future research can focus on this point.

CONCLUSION

In the field of bioinformatics, understanding protein secondary structure is very important for exploring diseases treatments. This study proposes a framework for predicting the protein secondary structure, consisting of multilevel features extraction, multistage combination classifiers, and multilevel samples discriminating module. In the multilevel features extraction module, we design a different backbone network to extract the features of the multilevel (easy and hard levels in this study) from the original data. In the multistage combination classifiers module, we design two classifiers to deal with samples with different difficulty levels, respectively. Finally, in the multilevel samples discriminating module, we design a measurement standard based on the Dirichlet distribution and information entropy to assign suitable samples to different classifiers (multistage combination classifiers) with different levels. The first classifier is used to learn and classify the easier samples and filter them out, avoiding being sent to the second classifier. Further, the remaining harder samples will be sent to the second classifier. We compute the loss value of the two classifiers. Consequently, the loss value of the harder samples will be accumulated and will always be greater than the easier ones. This method can induce the model to pay more attention to harder samples and improve the classification performance. The experimental results on the publicly available benchmark CB513 dataset show the superior performance of the proposed method.

However, the experimental results show that the current multilevel samples discriminating the module in this study are not designed well, which limits the performance of our framework. Herein, the related experiments show that if the multilevel samples discriminating module is designed well, our framework can obtain state-of-the-art performance. Besides, the depth of our network and the number of classifier branches also can be further increased to raise the performance of our framework. Hence, future work can focus on designing a more effective multilevel samples discriminating module and designing the deeper network as well as the more classifier branches to further improve the model performance.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization: XZ, LZ, and BJ; methodology, XZ, YL, LZ and YW; software, XZ; validation, YL, YW and LF; investigation, HZ, YW and LF; writing—original draft preparation, XZ and HZ; writing—review and editing, LZ, YL, BJ and LF; supervision, BJ; funding acquisition, BJ. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant no. 61772110) and the Key Program of Liaoning Traditional Chinese Medicine Administration (Grant no. LNZYXZK201910).

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi:10.1093/nar/25.17.3389

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). "Learning Imbal-Anced Datasets with Label Distribution-Aware Margin Loss," in *Advances in Neural Information Processing Systems*, 1565–1576.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic Minority Over-sampling Technique. *J. Artif. Intelligence Res.* 16, 321–357. doi:10.1613/jair.953

Cuff, J. A., and Barton, G. J. (2000). Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction. *Proteins* 40, 502–511. doi:10.1002/1097-0134(20000815)40:3<502:AID-PROT170>3.0.CO;2-Q

Cuff, J. A., and Barton, G. J. (1999). Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction. *Proteins* 34, 508–519. doi:10.1002/(SICI)1097-0134(19990301)34:4<508:AID-PROT10>3.0.CO;2-4

Cui, Y., Jia, M., Lin, T. Y., Song, Y., and Belongie, S. (2019). "Class-balanced Loss Based on Effective Number of Samples," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9268–9277. doi:10.1109/cvpr.2019.00949

Dempster, A. P. (2008). "A Generalization of Bayesian Inference," in *Classic Works of the Dempster-Shafer Theory of Belief Functions* (Berlin, Germany: Springer), 73–104. doi:10.1007/978-3-540-44792-4_4

Drori, I., Dwivedi, I., Shrestha, P., Wan, J., Wang, Y., He, Y., et al. (2018). *High Quality Prediction of Protein Q8 Secondary Structure by Diverse Neural Network Architectures*. arXiv [Preprint]. arXiv:1811.07143.

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015). JPred4: a Protein Secondary Structure Prediction Server. *Nucleic Acids Res.* 43, W389–W394. doi:10.1093/nar/gkv332

Duggal, R., Freitas, S., Dhamnani, S., Chau, D. H., and Sun, J. (2020). *ELF: An Early-Exiting Framework for Long-Tailed Classification*. arXiv [Preprint]. arXiv: 2006.11979.

Fang, C., Li, Z., Xu, D., and Shang, Y. (2020). MUFold-SSW: a New Web Server for Predicting Protein Secondary Structures, Torsion Angles and Turns. *Bioinformatics* 36, 1293–1295. doi:10.1093/bioinformatics/btz712

Fang, C., Shang, Y., and Xu, D. (2018). MUFOLD-SS: New Deep Inception-Inside-Inception Networks for Protein Secondary Structure Prediction. *Proteins* 86, 592–598. doi:10.1002/prot.25487

Feng, F. L., Chen, H. M., He, X. L., Ding, J., Sun, M., and Chua, T. S. (2019). *Enhancing Stock Movement Prediction with Adversarial Training*. California: IJCAI, 5843–5849.

Guo, Z., Hou, J., and Cheng, J. (2021). DNSS2 : Improved Ab Initio Protein Secondary Structure Prediction Using Advanced Deep Learning Architectures. *Proteins* 89, 207–217. doi:10.1002/prot.26007

Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2019). Improving Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility and Contact Numbers by Using Predicted Contact Maps and an Ensemble of Recurrent and Residual Convolutional Neural Networks. *Bioinformatics* 35, 2403–2410. doi:10.1093/bioinformatics/bty1006

Heffernan, R., Yang, Y., Paliwal, K., and Zhou, Y. (2017). Capturing Non-local Interactions by Long Short-Term Memory Bidirectional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers and Solvent Accessibility. *Bioinformatics* 33, 2842–2849. doi:10.1093/bioinformatics/btx218

Huang, Y., Chen, W., Dotson, D. L., Beckstein, O., and Shen, J. (2016). Mechanism of Ph-dependent Activation of the Sodium-Proton Antiporter Nhaa. *Nat. Commun.* 7, 12940. doi:10.1038/ncomms12940

Jens, M., Michael, M., Anita, Z., and Felix, S. (2001). Generation and Evaluation of Dimension-Reduced Amino Acid Parameter Representations by Artificial Neural Networks. *J. Mol. Model.* 7 (9), 360–369.

Jones, D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices 1 1Edited by G. Von Heijne. *J. Mol. Biol.* 292 (2), 195–202. doi:10.1006/jmbi.1999.3091

Josang, A. (2016). *Subjective Logic: A Formalism for Reasoning under Uncertainty*. Berlin, Germany: Springer. 978-3-319-42337-1.

Kabsch, W., and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22, 2577–2637. doi:10.1002/bip.360221211

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., et al. (2012). Template-based Protein Structure Modeling Using the Raptorx Web Server. *Nat. Protoc.* 7, 1511–1522. doi:10.1038/nprot.2012.085

Li, X., Zhong, C.-Q., Wu, R., Xu, X., Yang, Z.-H., Cai, S., et al. (2021). RIP1-dependent Linear and Nonlinear Recruitments of Caspase-8 and RIP3 Respectively to Necrosome Specify Distinct Cell Death Outcomes. *Protein Cell* 12, 858–876. doi:10.1007/s13238-020-00810-x

Li, Z., and Yu, Y. (2016). *Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks*. arXiv [Preprint]. arXiv:1604.07176.

Lyu, Z., Wang, Z., Luo, F., Shuai, J., and Huang, Y. (2021). Protein Secondary Structure Prediction with a Reductive Deep Learning Method. *Front. Bioeng. Biotechnol.* 9, 687426. doi:10.3389/fbioe.2021.687426

Minlong, P., Zhang, Q., Xing, X. Y., Gui, T., Huang, X. J., Jiang, Y. G., et al. (2019). Trainable Undersampling for Class-Imbalance Learning. *Proc. AAAI Conf. Artif. Intelligence* 33, 4707–4714.

Myers, J. K., and Oas, T. G. (2001). Preorganized Secondary Structure as an Important Determinant of Fast Protein Folding. *Nat. Struct. Biol.* 8, 552–558. doi:10.1038/88626

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proc. Natl. Acad. Sci.* 37, 205–211. doi:10.1073/pnas.37.4.205

Quan, L., Lv, Q., and Zhang, Y. (2016). STRUM: Structure-Based Prediction of Protein Stability Changes upon Single-point Mutation. *Bioinformatics* 32 (19), 2936–2946. doi:10.1093/bioinformatics/btw361

Shapovalov, M., Dunbrack, R. L., Jr, and Vucetic, S. (2020). Multifaceted Analysis of Training and Testing Convolutional Neural Networks for Protein Secondary Structure Prediction. *PLOS ONE* 15 (5), e0232528. doi:10.1371/journal.pone.0232528

Uddin, M. R., Mahbub, S., Rahman, M. S., and Bayzid, M. S. (2020). Saint: Self-Attention Augmented Inception-Inside-Inception Network Improves Protein Secondary Structure Prediction. *Bioinformatics* 36, 4599–4608. doi:10.1093/bioinformatics/btaa531

Wang, G., and Dunbrack, R. L., Jr. (2003). Pisces: a Protein Sequence Culling Server. *Bioinformatics* 19, 1589–1591. doi:10.1093/bioinformatics/btg224

Wang, S., Peng, J., Ma, J., and Xu, J. (2016). Protein Secondary Structure Prediction Using Deep Convolutional Neural fields. *Sci. Rep.* 6, 1–11. doi:10.1038/srep18962

Zhang, B., Li, J., and Lü, Q. (2018). Prediction of 8-state Protein Secondary Structures by a Novel Deep Learning Architecture. *BMC Bioinformatics* 19, 293. doi:10.1186/s12859-018-2280-5

Zhang, Y. (2008). I-tasser Server for Protein 3D Structure Prediction. *BMC Bioinformatics* 9, 40. doi:10.1186/1471-2105-9-40

Zhou, J., and Troyanskaya, O. (2014). "Deep Supervised and Convolutionalgenerative Stochastic Network for Protein Secondary Structure Prediction," in International Conference on Machine Learning (Beijing: PMLR), 745–753.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership